Spatio-temporal Event Detection and Forecasting in Social Media

Liang Zhao

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science

Chang-Tien Lu, Chair
Ing-Ray Chen
Jiangzhuo Chen
Narendran Ramakrishnan
Jieping Ye

June 29, 2016
Falls Church, Virginia

Keywords: event detection, event forecasting, social media

Spatio-temporal Event Detection and Forecasting in Social Media

Liang Zhao

(ABSTRACT)

Nowadays, knowledge discovery on social media is attracting growing interest. Social media has become more than a communication tool, effectively functioning as a social sensor for our society.

This dissertation focuses on the development of methods for social media-based spatiotemporal event detection and forecasting for a variety of event topics and assumptions. Five methods are proposed, namely dynamic query expansion for event detection, a generative framework for event forecasting, multi-task learning for spatiotemporal event forecasting, multi-source spatiotemporal event forecasting, and deep learning based epidemic modeling for forecasting influenza outbreaks. For the first of these methods, existing solutions for spatiotemporal event detection are mostly supervised and lack the flexibility to handle the dynamic keywords used in social media. The contributions of this work are: (1) Develop an unsupervised framework; (2) Design a novel dynamic query expansion method; and (3) Propose an innovative local modularity spatial scan algorithm.

For the second of these methods, traditional solutions are unable to capture the spatiotemporal context, model mixed-type observations, or utilize prior geographical knowledge. The contributions of this work include: (1) Propose a novel generative model for spatial event forecasting; (2) Design an effective algorithm for model parameter inference; and (3) Develop a new sequence likelihood calculation method. For the third method, traditional solutions cannot deal with spatial heterogeneity or handle the dynamics of social media data effectively. This work's contributions include: (1) Formulate a multi-task learning framework for event forecasting; (2) simultaneously model static and dynamic terms; and (3) Develop efficient parameter optimization algorithms.

For the fourth method, traditional multi-source solutions typically fail to consider the geographical hierarchy or cope with incomplete data blocks among different sources. The contributions here are: (1) Design a framework for event forecasting based on hierarchical multi-source indicators; (2) Propose a robust model for geo-hierarchical feature selection; and (3) Develop an efficient algorithm for model parameter optimization.

For the last method, existing work on epidemic modeling either cannot ensure timeliness, or cannot characterize the underlying epidemic propagation mechanisms. The contributions of this work include: (1) Propose a novel integrated framework for computational epidemiology and social media mining; (2) Develop a semi-supervised multilayer perceptron for mining epidemic features; and (3) Design an online training algorithm.

Spatio-temporal Event Detection and Forecasting in Social Media

Liang Zhao

(GENERAL AUDIENCE ABSTRACT)

Social media has experienced a rapid growth during the past decade. Millions of users of sites such as Twitter have been generating and sharing a wide variety of content including texts, images, and other metadata. In addition, social media can be treated as a social sensor that reflects different aspects of our society. Event analytics in social media have enormous significance for applications like disease surveillance, business intelligence, and disaster management. Social media data possesses a number of important characteristics including dynamics, heterogeneity, noisiness, timeliness, big volume, and network properties. These characteristics cause various new challenges and hence invoke many interesting research topics, which will be addressed here.

This dissertation focuses on the development of five novel methods for social media-based spatiotemporal event detection and forecasting. The first of these is a novel unsupervised approach for detecting the dynamic keywords of spatial events in targeted domains. This method has been deployed in a practical project for monitoring civil unrest events in several Latin American regions. The second builds on this by discovering the underlying development progress of events, jointly considering the structural contexts and spatiotemporal burstiness. The third seeks to forecast future events using social media data. The basic idea here is to search for subtle patterns in specific cities as indicators of ongoing or future events, where each pattern is defined as a burst of context features (keywords) that are relevant to a specific event. For instance, an initial expression of discontent gas price increases could actually be a potential precursor to a more general protest about government policies. Beyond social media data, in the fourth method proposed here, multiple data sources are leveraged to reflect different aspects of the society for event forecasting. This addresses several important problems, including the common phenomena that different sources may come from different geographical levels and have different available time periods. The fifth study is a novel flu forecasting method based on epidemics modeling and social media mining. A new framework is proposed to integrate prior knowledge of disease propagation mechanisms and real-time information from social media.

Acknowledgments

First and foremost, my deepest gratitude is to my advisor, Dr. Chang-Tien Lu, for his support during my 4 years of Ph.D. study. I have been incredibly fortunate to benefit from his guidance and invaluable insights on doing research. Beyond that, he has been a mentor and guide in my career and life. I will never forget his support and encouragement during the hardest part of my PhD years. It has been an absolute privilege to work with him for all these 4 years.

I am very thankful to all my committee members, who have provided not only important guidance for the completion of my dissertation, but also for my research work during our collaborations. Thanks go to Dr. Naren Ramakrishnan for offering his fantastic leadership of our group's major project, which has provided me with support throughout my 4-year research project. His deep expertise in writing and publishing has greatly benefited my research. Thanks also go to Dr. Jiangzhuo Chen and Dr. Ing-Ray Chen for providing insightful comments and valuable suggestions from my preliminary proposal, through my research defense, and for my final defense. Their knowledge of the field suggested new research directions that I might not otherwise have recognized. Special thanks go to Dr. Jieping Ye, who provided extensive expertise in sparse learning, that opened up new windows for my research. I am impressed and learnt a lot from his high efficiency and deep insights in scientific research.

I would like to especially thank the enormous support and continuous help I have received from Dr. Feng Chen over the last four years. His great personality and enormous passion for research have been always an excellent example for me.

I would like to express appreciation to my friends in the Spatial Data Management Laboratory: Ting Hua, Kaiqun Fu, Kevin Lu, Jing Dai, Fang Jin, Wei Wang (my roomate), Xuchao Zhang, Wei Wang (my colleague), Sathappan Muthiah, Taoran Ji, Tao Yu, Lei Zhang, Zhiqian Chen, and Weisheng Zhong, to name but a few - with whom I have shared unique moments thoughout this time. Thanks for being around and being there for me whenever I needed you all.. In the course of my Ph.D. study, they made the journey enjoyable with many happy memories.

Finally, I would like to dedicate this dissertation with special thanks to my wife, my parents, my uncle, and my parents-in-law. These amazing people have provided me with enormous support and love throughout this process. They are the reason why I am doing this.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, social media have became both a popular communication platform and a social sensor. Twitter is now one of the most popular microblogging services and social networks in the world [16] with 646 million users as of May 2015. Compared with traditional media, the posts in social media can be about any domain and any topic in the world, ranging from daily conversations to socially crucial issues. Thanks to the 140 character limitation of length, "timeliness" and "brevity" have become the most distinguishing features of tweets. This ensures the freshness of the Twitter posts, which usually beat traditional breaking news broadcasting media and make social media a promising information source for the most timely knowledge and news [2].

Social media has several salient characteristics: 1) Timeliness. Due to their brevity and the widespread use of mobile devices, tweets are commonly posted much faster than traditional media, where hours or even days are spent on compiling, proofreading, typesetting, and publishing; 2) Broad coverage of themes. Tweets cover almost every aspect of our lives, from everyday feelings to breaking news; and 3) Diverse channels for information dissemination. Twitter enables "retweeting" for fresh news cascading, "replying" for instant conversations, "hashtag" for theme tagging, and "friendship" for interest sharing. These characteristics make Twitter a highly valuable social sensor for tracking various interesting and crucial themes (e.g., crime and natural disasters), especially when the response times of traditional news outlets are too slow and cumbersome to provide useful information during emergencies and they may also be overseen by autocratic governments or threatened by criminal organizations [29].

This research focuses on the development of spatiotemporal event forecasting and detection methods, including dynamic query expansion for event detection, a generative framework for event forecasting, multi-task learning for spatiotemporal event forecasting, and deep learning based epidemics modeling for flu forecasting. These tasks have a wide array of applications, some of which are described below.

As its name suggests, spatiotempral event detection detects and tracks ongoing events as they occur for both time and location. This area is very close to the classic research on topic detection and tracking.

- Topic detection and tracking. A considerable body of work focuses on characterizing the general pattern of Twitter streams. The pattern is typically conceptualized as a mixture of "latent topics". For example, Blei and Lafferty aligned the proportion priors and distributions of latent topics over time [19]. Yang *et al.* proposed an efficient Twitter stream summarization approach that fits in limited memory [105]. Aiello *et al.* proposed an algorithm encompassing n-grams to detect trending topics [5]. Hong *et al.* analyzed the inter-relationships of multiple social media streams by considering both local topics and shared topics [52]. Mei and Zhai modeled latent topics through a mixture language model, and tracked the transitions among them [73]. However, because "latent topics" are typically extracted purely statistically, based on data without human prior knowledge, they do not necessarily have real-world meaning. Hence this thread of work is generally not appropriate for tracking targeted themes.

- Targeted-topic tracking. A thread of research in this area focuses on tracking targeted themes, such as earthquakes. The majority of this research adopts a classification framework to extract theme-related tweets based on solely contextual features [68, 90, 92], making it challenging to select an appropriate set of features. Li *et al.* proposed a generic framework for theme-related feature selection, although this approach is specially designed for scrawling two specific types of Twitter APIs and is not appropriate for the task attempted for this research [65]. A handful of methods have been proposed that take into account social relationships. Lin *et al.* implemented a probabilistic mixture model to characterize the temporal textual pattern and diffusion via friendship [67], while Ratkiewicz *et al.* applied a framework that was specifically designed to track the so-called "political astroturf" based on mentioning networks [87]. Li *et al.* suggested refining retrieved theme-related tweets by applying a classification-based method to predefined contextual and authorship-pivot features [64].

- Event detection in social media. There is a large amount of work on event detection in social media. Event detection methods typically utilize supervised (e.g., classifier) or unsupervised (e.g., graph clustering) frameworks to extract tweet subsets related to potential events that can be formalized as spatial burstiness [92, 109], temporal burstiness [4, 100], or spatiotemporal burstiness [63]. This thread of work has a different goal from our research topic, as it detects the emergence instead of the evolution of events, whereas our research focuses on continuously tracking the evolutionary dynamics of a theme.

- Query expansion in social media. Query expansion is a process that reformulates a seed query in order to improve the coverage and accuracy of information retrieval [26]. To improve the performance of retrieval in Twitter, a new thread of work utilizes

query expansion to dynamically expand keywords [10], retrieve tweets [71], and discover events [109]. The expanded keywords are typically extracted by exploring their co-occurrence with the user-specified initial query in a textual content, but information diffusion through social network has not yet been comprehensively explored.

- Event forecasting in social media. Most research in this area focuses on temporal events and ignores any underlying geographical information, generally looking at events such as forecasting the results of elections [78, 96], stock market movements [21], disease outbreaks [2, 89], box office ticket sales [9, 110], and crime [99]. These works can be grouped into three main categorizes: 1) Linear regression models, where simple features, such as tweet volumes, are utilized to predict the occurrence time of future events [9, 21, 49, 78]; 2) Nonlinear models, where more sophisticated features such as topic-related keywords are used as the input to build forecasting models using existing methods such as support vector machines or LASSO [89, 99]; and 3) Time series-based methods, where methods such as autoregressive models are used to model the temporal evolution of event-related indicators (e.g., tweet volume) [2]. However, few existing approaches provide true spatiotemporal resolution for the predicted events. In [41], Gerber utilized a logistic regression model for spatiotemporal events forecasting using topic-related tweet volumes as features. Wang et al. [98] developed a spatiotemporal generalized additive model to characterize and predict spatio-temporal criminal incidents, but their model requires the demographic data. Ramakrishnan et al. [86] built separate LASSO models for different locations to predict the occurrence of civil unrest events. The group at Virginia Tech [55, 86, 109] also designed a new query expansion method to expand both keywords and key tweets by considering both semantic and social network relationships, and used the burstiness of key tweets to predict civil unrest events. Zhao et al. [110] designed a new predictive model based on a topic model that jointly characterizes the temporal evolution in both semantics and geographical burstiness of social media content.

1.1 Research Issues

This research aims to investigate and develop social media based techniques for spatiotemporal event forecasting that are both efficient and effective. The major research issues are described in this section:

1.1.1 Unsupervised Spatial Event Detection in Targeted Topic

Twitter has become a popular data source for use when monitoring and detecting events. Targeted domains such as crime, elections, and social unrest require the creation of algorithms capable of detecting events that are pertinent to these domains. Due to the un-

structured language, short-length messages, dynamics, and heterogeneity that are typical of Twitter data streams, it is technically difficult and labor-intensive to develop and maintain supervised learning systems. This dissertation presents a novel unsupervised approach for detecting spatial events in targeted domains and illustrates this approach using a specific domain, viz. civil unrest modeling. Given a targeted domain, a dynamic query expansion algorithm is proposed to iteratively expand domain-related terms, and generate a tweet homogeneous graph. An anomaly identification method is utilized to detect spatial events over this graph by jointly maximizing local modularity and spatial scan statistics. Extensive experiments conducted in 10 Latin American countries demonstrate the effectiveness of the proposed approach.

1.1.2 A Generative Framework for Spatiotemporal Event Forecasting

Event forecasting based on social media data streams is an significant problem. The great majority of the existing approaches focus on forecasting temporal events (such as elections and sports) but are unable to forecast spatiotemporal events such as civil unrest and influenza outbreaks, which is a much more challenging task. To achieve spatiotemporal event forecasting, the time-evolving spatial features and their underlying correlations need to be considered and characterized. Here, batch and online approaches are applied for spatiotemporal event forecasting in social media such as Twitter. The proposed models characterize the underlying development of future events by jointly modeling the structural contexts and spatiotemporal burstiness based on different strategies. Both batch and online-based inference algorithms are developed to optimize the model parameters. Utilizing the trained model, the alignment likelihood of tweet sequences is calculated by dynamic programming. Extensive experimental evaluations on two different domains demonstrate the effectiveness of the new approach.

1.1.3 Multi-Task Learning for Spatio-Temporal Event Forecasting

Spatial event forecasting from social media, while important, suffers from a number of critical challenges, particularly due to the dynamic patterns of its features (keywords) and its geographic heterogeneity (e.g., spatial correlations, imbalanced samples, and different populations in different locations). Most existing approaches (e.g., LASSO regression, dynamic query expansion, and burst detection) address some, but not all, of these challenges, but not all of them. This research proposes a novel multi-task learning framework that is designed to concurrently address all these challenges. Specifically, given a collection of locations (e.g., cities), forecasting models are constructed for all the locations simultaneously by extracting and utilizing appropriate shared information that effectively increases the sample size for each location, thus improving the forecasting performance. Static features derived from a prede-

fined vocabulary by domain experts and dynamic features generated from dynamic query expansion are combined in a multi-task feature learning framework and different strategies investigated to balance homogeneity and diversity between the static and dynamic terms. The resulting algorithms based on Iterative Group Hard Thresholding are both efficient and effective for model training and prediction. Extensive experimental evaluations on Twitter data from four different countries in Latin America are presented that demonstrate the effectiveness of this new approach.

1.1.4 Multi-source Feature Learning for Spatial Event Forecasting

Forecasting significant societal events is an interesting and challenging problem as it must take into consideration multiple aspects of a society, including its economics, politics, and culture. Traditional forecasting methods based on a single data source find it hard to cover all these aspects comprehensively, thus limiting model performance. Multi-source event forecasting has proven promising but still suffers from several challenges, including 1) geographical hierarchies in multi-source data features, 2) missing values, and 3) characterization of structured feature sparsity. This research proposes a novel feature learning model that concurrently addresses all the above challenges. Specifically, multi-source data from different geographical levels, is applied to a new forecasting model by characterizing the lower-level features' dependence on higher-level features. To handle the correlations amidst structured feature sets and deal with missing values among the coupled features, a novel feature learning model is proposed that is based on an Nth-order strong hierarchy and fused-overlapping group Lasso.

1.1.5 Deep Learning Based Epidemic Modeling for Flu Forecasting

Infectious disease epidemics such as influenza and Ebola pose a serious threat to global public health. It is crucial to characterize the disease and the evolution of the ongoing epidemic rapidly, efficiently, and accurately. Computational epidemiology can model the disease progress and underlying contact network, but suffers from a lack of real-time and fine-grained surveillance data. Social media, on the other hand, provides timely and detailed disease surveillance, but is insensible to the underlying contact network and disease model. This research proposes a novel semi-supervised deep learning framework that integrates the strengths of computational epidemiology and social media mining techniques. Specifically, the new framework learns social media users' health states and intervention actions in real time, which are regularized by the underlying disease model and contact network. Conversely, the learned knowledge from social media can be fed into the computational epidemic model to improve the efficiency and accuracy of disease diffusion modeling. An online optimization algorithm to substantialize the above interactive learning process iteratively and

achieve a consistent stage of the integration. The extensive experimental results presented here demonstrate that the new approach is indeed capable of effectively characterizing spatiotemporal disease diffusion, outperforming competing methods by a substantial margin on multiple metrics.

1.2 Contributions

The major contributions of the research presented here can be stated as follows:

Unsupervised Spatial Event Detection in Targeted Topic

- **Development of an unsupervised framework:** A novel unsupervised approach for targeted domain spatial event detection in Twitter is presented here. The new method requires no intensive human labor such as training set labeling.
- **Design of a novel dynamic query expansion (DQE) method:** Given a targeted domain, DQE dynamically generates a set of domain-related key terms via a Twitter heterogeneous information network. The key terms are exhaustively extracted and then weighted appropriately based on DQE's iterative process.
- **An innovative local modularity spatial scan (LMSS) algorithm.** Based on a graph formed using key terms from DQE, LMSS jointly maximizes the local modularity and spatial scan statistics in order to distinguish events by taking into account both their semantic similarities and geographical proximities.

A Generative Framework for Spatiotemporal Event Forecasting

- **A novel generative framework for spatial event forecasting.** For spatial event forecasting in Twitter, an enhanced hidden Markov model (HMM) is presented that characterizes the transitional process of event development by jointly considering the time-evolving context and space-time burstiness of Twitter streams.
- **Effective batch and online algorithms for model parameter inference.** The model inference is formalized as the maximization of a posterior that is analytically tractable. Both EM-based algorithm and stochastic-EM parameter optimization algorithms are proposed to solve this problem effectively and efficiently.
- **A new sequence likelihood calculation method.** To handle the noisy nature of tweet content, words that are exclusive to a single event are identified by a language model that is optimized by a dynamic programming algorithm to achieve accurate sequence likelihood calculation.

- **An online parameter optimization algorithm.** To supplement existing methods, a new online parameter optimization algorithm was developed for this research. It not only reduces the training expense on large datasets but also improves the timeliness of the model in social media streams.

Multi-Task Learning for Spatio-Temporal Event Forecasting

1. **Formulation of a multi-task learning framework for event forecasting.** A new formulation for event forecasting for multiple cities in the same country is presented here as a multi-task learning problem. In the proposed model, event forecasting models are built for different cities simultaneously by restricting all cities to select a common set of features. Both penalized and constrained MTL formulations, which use different strategies to control the common set of features selected, are explored.

2. **Concurrent modeling of static and dynamic terms.** Existing models (LASSO and DQE) use different but complementary information; LASSO uses static terms, while DQE identifies dynamic terms. The MTL formulations proposed here make use of both types of information by integrating the strengths of LASSO (a supervised approach) and DQE (an unsupervised approach). To the best of our knowledge, there is little prior work that combines supervised and unsupervised approaches for event forecasting.

3. **Development of efficient algorithms.** Here, both convex and non-convex optimization formulations are explored. For convex problems, proximal methods are used, e.g., FISTA [13], as these have been shown to be efficient for solving sparse and multi-task learning problems. For non-convex problems, an iterative Group Hard Thresholding (IGHT) [20] framework is applied, as this is guaranteed to converge to a local solution.

Multi-source Feature Learning for Spatial Event Forecasting

- **Design a framework for event forecasting based on hierarchical multi-source indicators**. A generic framework is proposed for spatial event forecasting that utilizes hierarchically topological multiple data sources and is based on a generalized multi-level model. A number of classic approaches in the related research in this area are shown to be special cases of the new model.

- **Propose a robust model for geo-hierarchical feature selection**. To model the inherent structure in geo-hierarchical features across multiple data sources, an N-level interactive group Lasso is selected based on its strong hierarchy. To handle interactions among missing values, the proposed model adopts a multitask framework that is capable of learning the shared information among the tasks corresponding to all the missing patterns.

- **Develop an efficient algorithm for model parameter optimization.** To learn the proposed model, a constrained overlapping group Lasso problem needs to be solved, which is technically challenging. By developing an algorithm based on the alternating direction method of multipliers (ADMM) and introducing auxiliary variables, a globally optimal solution to this problem will be guaranteed.
- **Conduct extensive experiments for performance evaluations.** The proposed method was evaluated on 10 different datasets in two domains: forecasting civil unrest in Latin America and influenza outbreaks in the United States. The results demonstrate that the proposed approach runs efficiently and consistently outperforms the best of the existing methods along multiple metrics.

Deep Learning Based Epidemic Modeling for Flu Forecasting

- **Proposing a novel integrated framework for computational epidemiology and social media mining**: The existing approaches from computational epidemiology and social media mining focus on different but complementary aspects, with the former focusing on modeling the underlying mechanisms of disease diffusion while the latter provides timely and detailed disease surveillance. The new SimNest framework utilizes both types of information by integrating their strengths.
- **Developing a semi-supervised multilayer perceptron (MLP) for mining epidemic features**: To achieve deep integration,unsupervised pattern constraints derived from an epidemic disease progress model are enforced for the supervised classification. Using this semi-supervised strategy, the sparsity of labeled data can be solved.
- **Designing an online training algorithm**: To minimize the inconsistencies between Twitter space and the simulated world, model parameters can be iteratively optimized via an online algorithm. This algorithm ingests the social media data streams and updates the model parameters in real time, which not only reduces the cost of retraining but also ensures the timeliness of the model.

1.3 Organization of the Dissertation

The remainder of this dissertation is organized as follows. Chapter 2 defines the unsupervised spatiotemporal event detection framework, derives its major model, designs a new spatial clustering method, and presents the experimental results and discussion. Chapter 3 describes the proposed generative framework for event forecasting and presents an effective parameter inference algorithm, and the event forecasting method based on the proposed model. Chapter 4 presents the multitask formulation of event forecasting tasks, and proposes several models based on the new multitask framework, and then proposes an effective algorithm for

parameter optimization. Chapter 5 develops a social media-embedded epidemic modeling framework, and then presents an online algorithm for the model optimization. Chapter 6 summarizes the work carried out, lists the associated publications, and suggests directions for future research.

Chapter 2

Dynamic Query Expansion for Event Detection

This chapter presents a novel unsupervised approach for detecting spatial events in targeted domains and illustrate this approach using one specific domain, viz. civil unrest modeling. First, the introduction of this chapter is presented in Section 2.1, then the related work is summarized in Section 2.2.1. Section 2.2.2 introduces the problem of detecting targeted events, and Section 2.2.3 proposes the detailed methods to be used in our solution, together with a theoretical analysis. Extensive experiments along with comparisons to existing popular event detection methods and a case study are presented in Section 2.3. Finally, this chapter concludes with a summary of the study in Section 2.4.

2.1 Introduction

Microblogs such as Twitter and Weibo are experiencing an explosive level of growth. Millions of worldwide microblog users broadcast their daily observations on an enormous variety of domains, e.g., crime, sports, and politics. Traditional media, in contrast, is monopolized by closed groups, and on occasion may even be under threat from criminal organizations in localities suffering from conflicts and high crime rates [69]. When a social event occurs, it usually takes hours or even days to be reported by traditional media, which is why social media like Twitter have come to play a major role as a real-time information platform for social events [95, 103]. Beyond items of public interest, event-related microblogs can provide highly detailed and timely information for those interested in public safety, homeland security, and financial stability. Figure 2.1 depicts event hotspots related to the protests on September 27th, 2012 in Mexico. Based on tweets posted on that day, the new approach proposed here automatically and immediately identified these events, some of which were not reported by traditional media until several days later.

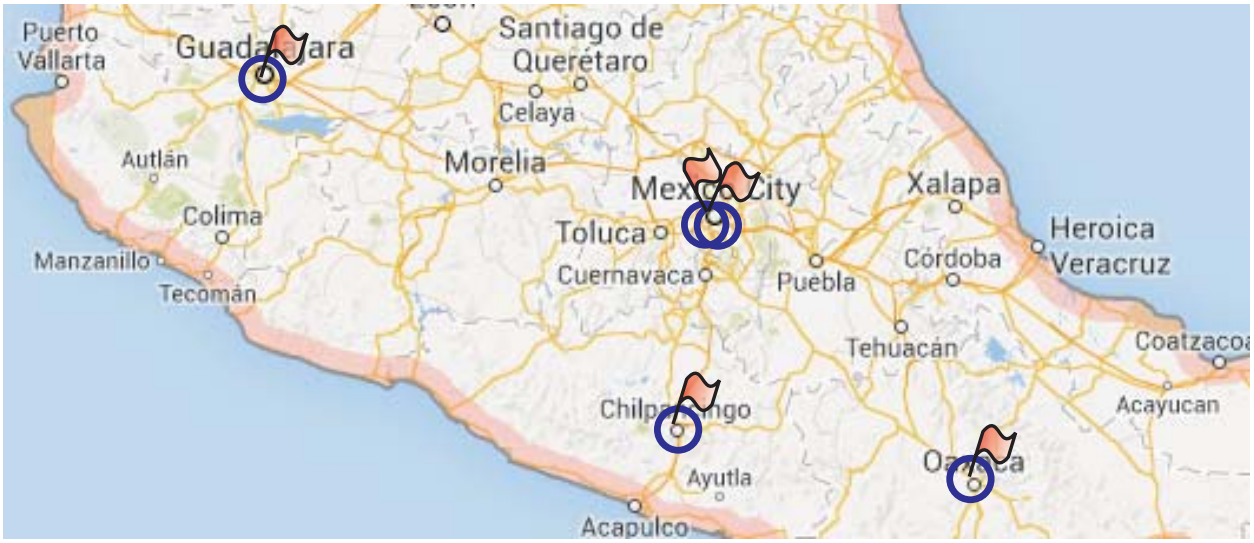


Figure 2.1: A map of civil unrest event hotspots on September 27th, 2012 pertaining to labor reform and other issues. Flags denote the ground-truth events reported by authorities. Circles denote the events detected by our method.

Although identifying events from news reports has been well studied [60], analyzing tweets to reveal event information requires more sophisticated techniques. Tweets are written in unstructured language and often contain typos, non-standard acronyms, and spam. In addition to the textual content, Twitter data forms a heterogeneous information network where users, tweets, and hashtags have mutual relationships. These features of Twitter data pose a challenge for event detection methods developed for traditional media. Although there has been a considerable body of work on event detection in Twitter, most of the work published has targeted events of *general interest*. Methods for *general interest* events typically focus on the "hotness" of events but are not sufficient for tracking events in specific domains. It is of high social significance to continuously and closely monitor crucial domains such as crime [64], earthquakes [90], civil unrest [86], and disease outbreaks [92]. Existing methods in event detection suffer from the following shortcomings: 1) their restricted ability to model heterogeneity and network properties of Twitter data. Existing methods typically treat Twitter data as a set of plain textual documents. However, "tweet", "word", "hashtag", and "user" are of different entity types. For example, a "user" can post a "tweet", "tweets" can be tagged by a "hashtag" and a "tweet" can reply to another "tweet". In general, these heterogeneous relationships and properties are not effectively harnessed by existing methods; 2) their limited ability to handle the dynamic properties of Twitter data. Existing methods treat fixed keywords as features for classifying tweets. However, the expression in tweets dynamically evolves, which makes the use of fixed features and historical training sets inappropriate. For example, the most significant Twitter keyword for the Mexican protests in Aug 2012 was "#YoSoy132" (i.e., the hashtag of an organization protesting against electoral fraud), alluding to the protests against the Mexican presidential election, but "#CNTE"

(i.e., the hashtag stands for the national teacher's association of Mexico) had become the most popular term by the beginning of 2013 due to the movements against the Mexican education reform; and 3) their inability to jointly model the semantic similarities and geographical proximities of events. Existing methods generally cannot differentiate between multiple events that occur simultaneously in the same location. For instance, in Mexico City, from Jan 30th, 2011 to Dec 31th, 2012, there were a total of 116 civil unrest events on 83 dates, of which 25 dates involved multiple events. On Sep 27, 2012, two different protests occurred in Mexico City, organized separately by "#Yosoy132" and "City sanitation workers". Hence, without the capacity to distinguish events' semantic contexts, existing methods typically miss nearly 30% of the events occurring in Mexico City.

In this work, to address the above-mentioned issues, this chapter presents an unsupervised "targeted domain" spatial event detection method that can jointly handle the heterogeneity and dynamics of Twitter data. Our contributions are summarized as follows:

- **Development of an unsupervised framework:** this chapter presents a novel unsupervised approach for targeted domain spatial event detection in Twitter. Our method requires no intensive human labor such as training set labeling.
- **Design of a novel dynamic query expansion (DQE) method:** Given a targeted domain, DQE dynamically generates a set of domain-related key terms via a Twitter heterogeneous information network. The key terms are exhaustively extracted and then weighted appropriately based on DQE's iterative process.
- **An innovative local modularity spatial scan (LMSS) algorithm.** Based on a graph formed using key terms from DQE, LMSS jointly maximizes the local modularity and spatial scan statistics in order to distinguish events by taking into account both their semantic similarities and geographical proximities.
- **Extensive experimental evaluation and performance analysis.** Our method was extensively evaluated on Twitter data covering 10 Latin American countries. Comparisons with baselines and state-of-the-art methods demonstrated its effectiveness and efficiency.

2.2 Materials and Methods

2.2.1 Literature Review

Current microblog-based event detection methods can be classified into two categories: 1) *general-interest event detection*, and 2) *targeted-domain event detection*.

General-interest event detection. Methods under this category aim to detect emerging general-interest topics in the Twitter data stream, and typically apply unsupervised techniques such as topic modeling, burst detection, and clustering techniques. Yin et al. [106] developed geographic topic modeling techniques to detect topics clustered in local geographic regions, while Lappas et al. [63] proposed methods to discover bursts of terms in a specific spatial and temporal neighborhood. Weng et al. [100] applied wavelet analysis for noise filtering and then identified word groups with high correlations, each of which is returned as the indicator of an event. Adopting a different approach, Aggarwal and Subbian [4] developed an algorithm that captures the related signals by considering the tweets' content, network structural, and temporal information. Finally, Ritter et al. [88] suggested an NLP-based approach to general event extraction from twitter data.

Targeted-domain event detection. Methods under this category aim to detect events within a particular field, e.g., "earthquakes", "disease outbreaks", and "civil unrest". These methods generally rely on supervised learning techniques like the support vector machine (SVM). Human labor is required to label the subsets of tweets related to the targeted domain, and then clustering techniques are applied to identify the locations of the events. An example of this is a study by Sakaki et al. [90], who designed a classifier to extract earthquake-related tweets and then utilized Kalman filtering to detect the geographic regions where the earthquakes had occurred. For tracking disease activities, Signorini et al. [92] adopted an SVM classifier to extract tweets related to various types of disease, while Chakrabarti et al. [29] trained a modified Hidden Markov Model to learn the structure and vocabulary of sports-related tweets, which were then utilized to generate summaries of the sports events. Li et al. [64] trained a classifier to extract crime-related tweets, first sorting the tweets based on their importance, and then applying them to detect crime events.

2.2.2 Problem Formulation

Twitter data contains heterogeneous entities and multiple types of relationships, which can be formulated as a Twitter heterogeneous information network:

Definition 1. ***(Twitter Heterogeneous Information Network)*** *A* **Twitter heterogeneous information network** *is defined as an* undirected *graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W}, \mathcal{S})$*, where* $\mathcal{V} = \mathcal{T} \cup \mathcal{F}$*.* $\mathcal{T}$ *refers to a set of* **tweet nodes***, and* $\mathcal{F} = \mathcal{F}_1 \cup \cdots \cup \mathcal{F}_M$ *refers to other* M *types (e.g., term, user, and hashtag) of nodes, called* **feature nodes***.* $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ *represents the set of edges, which are all undirected. We denote the existence of an edge between two nodes* $v_i, v_j \in \mathcal{V}$ *by* $v_i \leftrightarrow v_j$*.* $\mathcal{W}$ *denotes the set of weights of nodes and edges.* $\mathcal{S} = \{l(v) | v \in \mathcal{T}\}$ *refers to a set of geographic locations of tweet nodes, where* $l(v) \in \mathcal{R}^2$ *represents a tuple consisting of the latitude and longitude of tweet node* v*. When* $M = 0$*,* $\mathcal{G}$ *reduces to a* **Twitter homogeneous information network** $\mathcal{G}_0$*.*

In addition to ***tweet nodes***, several other types of nodes are considered, including "term", "hashtag", "hyperlink", and "user", all of which are generally called ***feature nodes***. The

relationships between these types of nodes are denoted by the set of undirected edges $\mathcal{E}$, including *authorship* between user nodes and tweet nodes, *containment* between tweet nodes and term nodes, and *replying* between tweet nodes.

Definition 2. ***(Seed Query)*** *A* **seed query** *is defined as an initial set of semantically coherent feature nodes that characterize the concept of the targeted domain. A seed query is denoted as* $\mathcal{Q}_0 = \{(v_i, w(v_i)^{(0)})\}_{i=1}^{N}$*, where the feature node* v_i *is a* **seed query term** *whose weight* $w(v_i)^{(0)} \in R^+$ *reflects its relevance to the targeted domain. An* ***expanded query*** *is an extended set of weighted feature nodes that represent the semantic contexts of spatial events. Similar to seed query, an expanded query is denoted as* $\mathcal{Q} = \{(v_i, w(v_i))\}_{i=1}^{N'}$*, where* v_i *is called an* **expanded query term**.

All the *seed query terms* have corresponding edges denoting their semantic relevance. For example, given a seed query of the domain "civil unrest": {("protest", 1), ("march", 1), ("strike", 1), ("unrest", 1)}, an expanded query can be: {("#megamarcha", 0.1), ("#YoSoy132", 0.3), ("zocal",0.1), ("march",0.2), ("imposicin",0.1)}, which matches the news description: "A mega march against the imposition of PRI: YoSoy132 protestors arrived at El Zocalo."

Denote $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_P\}$ as a collection of time-indexed Twitter data, where $\mathcal{C}_p \in \mathcal{C}$ represents the subcollection of tweets posted between timestamps t_{p-1} and t_p. To achieve targeted domain spatial event detection, one needs to concentrate on domain-related tweets and detect the spatial burst signals based on them. The major tasks of the **targeted-domain spatial event detection** problem are defined as follows:

Task 1: Expanded Query Generation: Given $\mathcal{C}_p$ and a seed query $\mathcal{Q}_0$ of a targeted domain, ***expanded query generation*** is to generate the expanded query $\mathcal{Q}_p$ by expanding $\mathcal{Q}_0$ through the Twitter heterogeneous information network $\mathcal{G}$.

Task 2: Spatial Event Extraction: Given a targeted-domain related tweets subset extracted based on $\mathcal{Q}_p$ from $\mathcal{C}_p$, ***spatial events extraction*** is to automatically identify a set of spatial events, each of which is specified by geolocation, time, and related tweet nodes.

2.2.3 Dynamic Query Expansion

Specially designed for Twitter data, the dynamic query expansion (DQE) algorithm utilizes heterogeneous relationships (e.g., containment, authorship, and replying) extracted from the Twitter heterogeneous information network to expand the seed query. The leftmost component in Figure 2.2 shows the general framework of the DQE algorithm.

Calculation of Relevances to Targeted Domain

Given a seed query, we must first focus on generating an expanded query. Traditional query expansion methods generally expand the seed query by examining the terms' semantic or co-

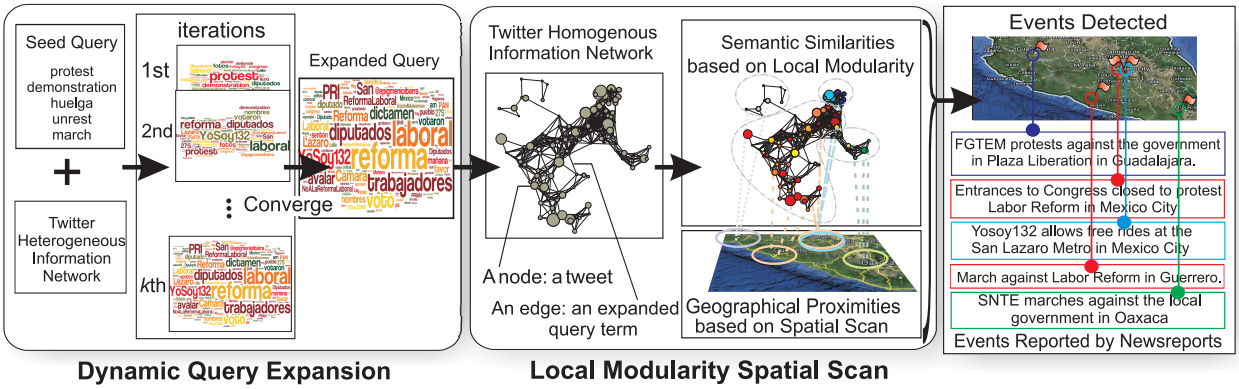


Figure 2.2: Flowchart of the proposed method.

occurrence relationships in textual documents. To further enhance the coverage of expanded query, recently Li et. al proposed to expand the query iteratively by taking into account the usefulness and coverage of the keywords in each iteration [65]. However, their method is not guaranteed to converge and is sensitive to the number of iterations user specified. Most existing query expansion methods are seriously challenged by the heterogeneity of Twitter data. *First, Twitter data contains multiple types of entities.* In addition to terms, entities such as "users", "hashtags", and "hyperlinks" are all important for revealing the implicit relevance between tweets. For example, a "keyplayer" (i.e., an important Twitter user in a particular domain of activity) in a particular domain will frequently post domain-related tweets, and thus the tweets and terms posted by him/her are likely to be domain-related. *Second, Twitter data contains heterogeneous relationships among multiple types of entities.* Social relationships in Twitter provide heuristics to associate tweets under a same domain. For example, tweets can have mutual social relationships such as "replying" or "replied". A tweet and the tweets replying to it will therefore generally fall into the same domain.

To overcome these challenges, a new dynamic query expansion algorithm is proposed that utilizes the heterogeneous relationships of Twitter heterogeneous information network. Referring to Definition 1, for any node $v_i \in \mathcal{V}$, its weight $0 \leq w(v_i) \in \mathcal{W}$ is defined as its relevance to the targeted domain. The nodes with higher weights are more relevant to the targeted domain. For example, "protest" and "#OccupyWallSt" are more relevant to the "civil unrest" domain than "love" and "#music", thus the weights of "protest" and "#OccupyWallSt" are higher. To simplify the notation, for any $V \subseteq \mathcal{V}$, the weights set $\{w(v_i)|v_i \in V\}$ is denoted as $w(V)$.

Heterogenous relationships among entities are ubiquitous and important in Twitter. Terms such as "protest" are deemed to be related to the "civil unrest" domain because they appear frequently in the set of domain-related tweets. Similarly, user "ESPN" is related to the "sports" domain because "ESPN" mainly posts tweets about sports; tweets tagged by the hashtag "#OccupyWallSt" are considered to be about "civil unrest". A tweet is typically deemed to be in the same domain as the one it replies to. Tweet nodes and feature nodes

generally exhibit a mutual reinforcing relationship. Given a set of feature nodes $F \subseteq \mathcal{F}$ and tweet nodes $T \subseteq \mathcal{T}$, if a feature node $v_i \in F$ has edges with many high-weight tweet nodes instead of low-weight ones, it should receive a large weight value. Then, if $v_j \in T$ has edges with many high-weight feature nodes, it should also be assigned a large weight value. It also follows that, if $v_j \in T$ has a replying relationship with a high-weight tweet node $v_k \subseteq T$, v_j should also receive a large weight value. The first of these relationships determines the weights of feature node set F while the second and third determine the weights of tweet node set T.

The operation to determine the weights of the nodes in F proceeds as follows:

$$w(F) = D_F \cdot A_{F,T} \cdot w(T), \tag{2.1}$$

where $w(F)$ and $w(T)$ denote the vector weights of F and T, respectively. $A_{F,T}$ denotes the adjacency matrix between F and T such that $[A_{F,T}]_{ij} = 1$ if $v_i \leftrightarrow v_j$, where $v_i \in F, v_j \in T$; $[A_{F,T}]_{ij} = 0$, otherwise. D_F is the inverse document frequency (IDF) [100] matrix of all the words in the vocabulary $\mathcal{V}$, which is a diagonal matrix such that $[D_F]_{ii}$ refers to the IDF of $v_i \in F$.

The operation to determine the weights of the nodes in T proceeds as follows:

$$w(T) = A'_{F,T} \cdot w(F) + \beta A_T \cdot w(T), \tag{2.2}$$

where β reflects the tradeoff between the influences of feature nodes and tweet nodes on the calculation of $w(T)$. A_T denotes the matrix of the replying relationship between tweet nodes such that $A_{T_{i,j}} = 1$ if $v_i \leftrightarrow v_j$, where $v_i, v_j \in T$; $A_{T_{i,j}} = 0$, otherwise. $A'_{F,T}$ is the transpose of matrix $A_{F,T}$.

DQE Algorithm Description

To generate an expanded query, above-mentioned operations are utilized via an iterative DQE algorithm, as shown in Table 2.1 Algorithm 1. The major issues of the algorithm implementation are described in the following.

Initialization. Suppose we are given a seed query $\mathcal{Q}_0 = \{(v_i, w(v_i)^{(0)})\}_{i=1}^{M}$ for the targeted domain. Denote $w(\cdot)^{(k)}$ as the weight(s) of the node(s) at the kth iteration. Denote $T_r^{(k)}$ as the set of domain-related tweet nodes at the kth iteration. To trigger the iterative operations, T_r^0 is initialized as the set of tweet nodes matching $\mathcal{Q}_0$. All the feature nodes having edges with nodes in T_r are potentially domain-related and thus can be used to initialize the feature node set $F \subseteq \mathcal{F}$. The initial tweet node set $T \subseteq \mathcal{T}$ consists of tweet nodes, each of which has edge(s) with at least one node in F. Naturally, $T_r^0 \subseteq T$. $w(T_r^0)^{(0)}$ is an all-one vector while $w(T - T_r^0)^{(0)}$ is a zero vector.

Stopping Criterion. For the kth iteration, tweet nodes in $T_r^{(k)}$ are compared to those in $T - T_r^{(k)}$ based on their weights. If $\forall v_i \in T_r^{(k)}$ and $\forall v_j \in T - T_r^{(k)}$: $w(v_i) \geq w(v_j)$, then the

Algorithm 1: Dynamic Query Expansion.

Input: Seed Query $\mathcal{Q}_0 = \{(v_i,\ w(v_i)^{(0)})\}_{i=1}^{M}$, Twitter sub-collection $\mathcal{C}_p$

Output: Expanded Query $\mathcal{Q}_p$

Initialize T, F, T_r^0, and $w(T)$

Set Φ via Equation 2.3 and 2.4

Set $k = 0$

repeat

  repeat

    $Swap\left(\min\left(w(T_r^k)^{(k)}\right), \max\left(w(T - T_r^k)^{(k)}\right)\right)$

    $\sigma = \min\left(w(T_r^k)^{(k)}\right) - \max\left(w(T - T_r^k)^{(k)}\right)$

  until $\sigma \geq 0$;

  $w(F)^{(k)} = D_F \cdot A_{F,T} \cdot w(T)^{(k-1)}$

  $w(T)^{(k)} = \Phi \cdot (A'_{F,T} \cdot w(F)^{(k)} + \beta A_T \cdot w(T)^{(k-1)})$

  $\sigma = \max\left(w(T - T_r^k)^{(k)}\right) - \min\left(w(T_r^k)^{(k)}\right)$

  $k = k + 1$

until $\sigma \leq 0$;

$w(F_r) = \{w(v_i)^{(k)} \in w(F)^{(k)} | v_i \in F_r \subseteq F\}$

$\mathcal{Q}_p = \{(v_i, w(v_i)) | v_i \in F_r, w(v_i) \in w(F_r)\}$.

Table 2.1: The algorithm of Dynamic Query Expansion

iterations will be terminated, as shown in Line 13. Otherwise, the lowest-weight node in $T_r^{(k)}$ will be exchanged with the highest-weight node in $T - T_r$ (denoted by the function "Swap" in Line 6) until $\forall v_i \in T_r^{(k)}$ and $\forall v_j \in T - T_r^{(k)}$: $w(v_i) \geq w(v_j)$, as shown in Line 8.

Generation of the Expanded Query. After the iterations are completed, the ultimate set of domain-related tweet nodes is $T_r^{(k)}$. Define a set F_r of feature nodes, each of which has edge(s) to at least one node of $T_r^{(k)}$. Due to $F_r \subseteq F$, the weights of the nodes in F_r have been calculated, as shown in Line 14, and eventually the expanded query Q^* is formed in Line 15.

Analysis of Convergence. Equations 2.1 and 2.2 are combined to capture the weight updating of T:

$$w(T)^{(k)} = E \cdot w(T)^{(k-1)}, \tag{2.3}$$

where the matrix E is a transition matrix (column-normalized by Φ) consisting of the relevances between any two tweet nodes in T:

$$E = \Phi \cdot (A'_{F,T} \cdot D_F \cdot A_{F,T} + \beta A_T), \tag{2.4}$$

where Φ normalizes $A'_{F,T} \cdot D_F \cdot A_{F,T} + \beta A_T$ by column so that the weights in $w(T)^{(k)}$ sum to a constant.

Formulate three facts introduced above: 1) $\forall v_i \in T$, $\exists v_j \in F$: $v_i \leftrightarrow v_j$, 2) $\forall v_j \in F$, $\exists v_k \in \{v|(v, w(v)) \in \mathcal{Q}_0\}$: $v_j \leftrightarrow v_k$, and 3) $\forall v_l, v_m \in \{v|(v, w(v)) \in \mathcal{Q}_0\}$: $v_l \leftrightarrow v_m$.

Therefore, we obtain $\forall v_a, v_b \in T$, $\exists v_c, v_f \in F$ and $\exists v_d, v_e \in \{v|(v, w(v)) \in \mathcal{Q}_0\}$: $v_a \leftrightarrow v_c \leftrightarrow v_d \leftrightarrow v_e \leftrightarrow v_f \leftrightarrow v_b$, which means any two nodes in T have a path connected to each other. Hence, E is irreducible because its corresponding graph formed by T is strongly connected [42].

The Markov chain associated with E is irreducible. In addition, its aperiodicity is guaranteed [48]. Therefore, this Markov chain is ergodic. Based on the stability theorem of Markov chains, the existence of a unique stationary distribution vector for this Markov chain is guaranteed [42], which means as k increases, $w(T)^{(k)}$ converges to $w(T)^*= \lim_{k\to\infty} E^k$. Therefore, the convergence is guaranteed.

2.2.4 Local Modularity Spatial Scan

We describe a local modularity spatial scan (LMSS) model that can be applied to extract spatial events, as illustrated in the corresponding component in Figure 2.2. Based on the tweet graph built with the expanded query, we first derive an optimization function for identifying anomalous subgraphs, and then apply this function to identify tweets related to latent spatial events.

Anomalous Subgraph Identification

The expanded query Q_p contains the feature nodes that are most relevant to the targeted domain. Q_p is utilized to retrieve the set of domain-related tweets T_{Q_p}, in which each tweet contains at least one of the expanded query terms. We need to extract tweet node sets $\{V_i\}_{i=1}^n$, where each $V_i \subseteq T_{Q_p}$ contains tweets related to a latent spatial event. This is typically solved by spatial clustering methods [64, 90]. However, if only the geographic proximities in clustering are considered, it is not possible to distinguish between discrete events when they occur in the same location.

To address this problem, the semantic similarities and geographical proximities of tweets are jointly considered based on the Twitter homogeneous information network. The event-related tweets need to be both semantically similar and geographically close. Specifically, by referring to Definition 1, a Twitter homogeneous information network $\mathcal{G}_0 = (\mathcal{V}_0, \mathcal{E}_0, \mathcal{W}_0, \mathcal{S}_0)$ can be built, where $\mathcal{V}_0 = T_{Q_p}$ denotes the node set, $\mathcal{S}_0 = \{l(v)|v \in \mathcal{V}_0\}$ stands for the tweet nodes' geographic locations, and $\mathcal{E}_0 = \mathcal{V}_0 \times \mathcal{V}_0$ represents the set of undirected edges. In addition, the weight set $\mathcal{W}_0 = w(\mathcal{E}_0)$ represents the semantic similarities among tweet nodes such that two tweets are semantically similar if they share expanded query terms. Mathematically, $w(\mathcal{E}_0) = A \cdot A^T$ where A is the adjacency matrix between T_{Q_q} and F_r. Since $\mathcal{E}_0$, $\mathcal{W}_0$, and $\mathcal{S}_0$ all depend on $\mathcal{V}_0$, for convenience, $\mathcal{G}_0 = (\mathcal{V}_0, \mathcal{E}_0, \mathcal{W}_0, \mathcal{S}_0)$ is denoted as $\mathcal{G}(\mathcal{V}_0)$. The graph $\mathcal{G}(V_1)$ is said to be a subgraph of graph $\mathcal{G}(V_2)$ if $V_1 \subseteq V_2$. Hence, in $\mathcal{G}(\mathcal{V}_0)$, the event-related tweets are deemed to compose a subgraph $G = \mathcal{G}(V \in \mathcal{V}_0)$ that

Algorithm 2: Local Modularity Spatial Scan.

Input: $\mathcal{G}_0 = (\mathcal{V}_0, \mathcal{E}_0, \mathcal{W}_0, \mathcal{S}_0)$
Output: $\Omega = \{\mathcal{G}(V_i)\}_{i=1}^{K}$, where $V_i \subseteq \mathcal{V}_0$
Initialize $\Omega = \emptyset$
for $s \in \mathcal{S}_0$ **do**
  $\mathcal{V}_s = \{v | v \in \mathcal{V}_0, l(v) \in \mathcal{S}_s\}$
  $V^* = \arg\max_{V \subseteq \mathcal{V}_s} f_l(\mathcal{G}(V))$, s.t. $\mathcal{G}(V)$ is connected
  repeat
    $V_r^* = \arg\max_{V_r \subseteq \mathcal{V}_s - V^*} f(\mathcal{G}(V^* \cup V_r))$, s.t. $\mathcal{G}(V_r)$ is connected
    $V^* = V^* \cup V_r^*$
    $p = \arg\min_{v \in V^*} H(l(v))$
    $q = \arg\max_{v \in \mathcal{V}_s - V^*} H(l(v))$
    if $H(p) < H(q)$ **then**
      $V^* = V^* \cup \{q\} - \{p\}$
      $\mathcal{V}_p = (\mathcal{V}_s - V^*) \cup \{p\} - \{q\}$
      $V^* = \arg\max_{V \subseteq \mathcal{V}_p} f(\mathcal{G}(V^* \cup V))$, s.t. $\mathcal{G}(V)$ is connected
  until $H(p) \geq H(q)$;
  Add $\mathcal{G}(V^*)$ to Ω
Check overlapping among subgraphs and update Ω
Randomization testing on subgraphs and update Ω

Table 2.2: The algorithm of Local Modularity Spatial Scan

satisfies two properties: 1) tweets in G connect via high-weight edges, and 2) tweets in G are geographically proximate with each other.

For the first property, local modularity [75] is adopted, which is a metric generally applied to measure the quality of a connected subgraph:

$$f_l(G) = L^G / L_N^G - L_{in}^G L_{out}^G / (L_N^G)^2, \tag{2.5}$$

where $L^G = L_{in}^G + L_{out}^G$. $f_l(G)$ is the local modularity of $G = \mathcal{G}(V \in \mathcal{V}_0)$, an arbitrary subgraph of $\mathcal{G}(\mathcal{V}_0)$, L_{in}^G refers to the sum of the weights of the edges in G, L_{out}^G denotes the sum of the weights of the edges that connect nodes in G and nodes outside G, and L_N^G represents the sum of the weights of the edges in the subgraph formed by the nodes in the geographical neighborhood of the nodes of V.

For the second property, Kulldorff proposes an effective metric to measure the geographical proximities of a spatial cluster, dubbed the Kulldorff statistic [59]. It is applied to measure the geographical proximities of the subgraph G:

$$f_s(G) = C \log \frac{C}{B} + (C_{all} - C) \log \frac{C_{all} - C}{B_{all} - B} - C_{all} \log \frac{C_{all}}{B_{all}}, \tag{2.6}$$

where C refers to the count of tweet nodes in $G = \mathcal{G}(V \in \mathcal{V}_0)$, B refers to the size of the set of tweet nodes $V_B = \{v|v \in \mathcal{V}_0, l(v) \in \{v_i|v_i \in V\}\}$, C_{all} denotes the count of tweet nodes in $\mathcal{V}_0$, and B_{all} represents the count of tweet nodes in $\mathcal{T}$.

Hence, Task 2 can be addressed by identifying the anomalous subgraphs that jointly maximize the preceding two quality metrics. This is formalized as a multi-objective optimization problem as follows:

$$\max_{V \subseteq \mathcal{V}': \mathcal{G}(V) \text{ is connected}} f(\mathcal{G}(V)) = f_l(\mathcal{G}(V)) + \lambda h_s(\mathcal{G}(V)), \tag{2.7}$$

where λ is a predefined parameter to balance the significance of the local modularity for semantic similarities and the Kulldorff statistics for spatial proximities. $\mathcal{V}'$ is an arbitrary subset of $\mathcal{V}_0$.

LMSS Algorithm

By exploring the linear-time subset scanning (LTSS) property of the Kulldorff statistic [77], this chapter proposes a fast approximate algorithm (Algorithm 2 in Table 2.2) that adopts a heuristic strategy to search for anomalous subgraphs that maximize $f(G)$ in Equation 2.7. The algorithm is elaborated as follows.

Anomalous Subgraphs Extraction (Line 2-4). Each distinct geographic location $s \in \mathcal{S}_0$ is considered as a candidate geographic center (Line 2). A tweet node set $\mathcal{V}_s \subseteq \mathcal{V}_0$ is first extracted with a corresponding set $\mathcal{S}_s \subseteq \mathcal{S}_0$ consisting of locations within a distance r of the center s (Line 3). In each $\mathcal{V}_s$, by applying a local modularity graph clustering algorithm [75], a subset V^* is found that has the maximum local modularity (Line 4).

Subgraph Refinement (Line 5-15). The proposed algorithm then finds a connected subgraph $V_r^* \subseteq \mathcal{V}_s - V^*$ that maximizes $f(\mathcal{G}(V^* \cup V))$ (Line 6), where V is a subset of $\mathcal{V}_s - V^*$. Then V^* is updated by merging it with V_r^* (Line 7). To achieve linear time subgraph scanning, Neill proposed a statistic priority function $H(s)$ for location s such that $H(s) = N_c/N_b$, where N_c and N_b are the numbers of tweet nodes on location s in the subgraph and in the whole graph, respectively [77]. If the minimum value of the statistical priorities of the locations of V^* is larger than the maximum value of those of $\mathcal{V}_s - V^*$ (Line 8-10), add $\mathcal{G}(V^*)$ into the graph list Ω (Line 15). Otherwise, exchange the minimum-value location of V^* with the maximum-value location of $\mathcal{V}_s - V^*$ (Line 11-12), and update V^* by finding a subgraph $V^* \cup V$ that maximizes $f(\mathcal{G}(V \cup V^*))$, where $V \subseteq V_p$ (Line 13).

Candidate Subgraph Set Pruning (Line 16-17). If there exist subgraphs $\Omega' = \{\mathcal{G}(V_i)\}_{i=1}^{K'} \subseteq \Omega$ sharing the same nodes, retain only the subgraph $\mathcal{G}(V) = \arg\max_{\mathcal{G}(V) \in \Omega'} f(\mathcal{G}(V))$ (Line 16). Then $f(G)$ of each subgraph $G \in \Omega$ is tested using randomization testing, and retain only the subgraphs with empirical p-values smaller than 0.05 (Line 17).

The LMSS algorithm exhibits several advantageous theoretical properties, as follows:

Theorem 1. *Algorithm 2 in Table 2.2 has the following theoretical properties: If $\lambda = 0$, it is guaranteed to return a local optimal solution that maximizes the local modularity score $f(G)$; If $\lambda = +\infty$, it is guaranteed to return a global optimal solution that maximizes the Kulldorff statistic $f_s(G)$.*

Proof. If $\lambda = 0$, then the solution of Line 4 will be returned as the final value of V^* for s, which proves the first property in Theorem 1. If $\lambda = +\infty$, then the Kulldorff statistic dominates $f(G)$. Line 6 searches for the set of tweet nodes $V_r^* \subseteq \mathcal{V}_s - V^*$ by maximizing $f(\mathcal{G}(V_r^* \cup V_r))$. Note that in this step, the set of locations of tweet nodes in V^* is fixed, and hence the factors C, B and B_{all} are fixed. Given that $\lambda = +\infty$ and $f_s(G)$ is a homogeneous function of the count C, the optimal solution V_r^* is identical to $\mathcal{V}_s - V^*$. Recall the basic idea of the LTSS property [77]: the subset of geographic locations that maximizes the Kulldorff statistic can be found by ranking the locations according to the priority function $H(s)$, and then searching over groups consisting of k locations with highest priority values. It can be readily proved that by solving the objective function in Line 13, the resulting V^* will be the connected subgraph consisting of the locations with the highest priority values. Hence, the Kulldorff statistic $f_s(G)$ will be maximized. □

2.2.5 Time Complexity Analysis

The time complexity of DQE is $O(l \cdot (|\mathcal{F}| \cdot n_{E_{TF}} + |\mathcal{T}| (n_{E_{TF}} + n_{E_{TT}})))$, where $\mathrm{n}_{E_{TF}} \ll |\mathcal{F}|$ is the average number of connections between a tweet node and feature nodes, $\mathrm{n}_{E_{TT}} \ll |\mathcal{T}|$ is the average number of connections from a tweet node to other tweet nodes, and l is the number of the iterations of DQE. Typically, $l \leq 10$.

The time complexity of LMSS is $O(\sum_s |\mathcal{S}_s| \log |\mathcal{S}_s| + |V_r^*| \cdot \sum_s |\mathcal{V}_s|^2) = O\left(|V_r^*| \sum_s |\mathcal{V}_s|^2\right)$, where $\sum_s |\mathcal{S}_s| \log |\mathcal{S}_s|$ corresponds to the solving of the objective function in Line 6 of Table 2.2 Algorithm 2 while $|V_r^*| \cdot \sum_s |\mathcal{V}_s|^2$ corresponds to the local modularity calculation. $|\mathcal{S}_s| \leq |\mathcal{V}_s|$ and $|\mathcal{V}_s| < |\mathcal{T}'| \ll |\mathcal{T}|$, where $\mathcal{T}'$ is the set of the tweet nodes with weights higher than 0.

By summing up these two parts, which correspond to DQE and LMSS, respectively, the overall time complexity is $O(l \cdot (|\mathcal{F}| \cdot n_{E_{TF}} + |\mathcal{T}| (n_{E_{TF}} + n_{E_{TT}})) + |V_r^*| \sum_s |\mathcal{V}_s|^2)$.

2.3 Results

In this section, the empirical evaluations of the performance of our approach, DQE+LMSS, are presented. By comparing the results with those obtained using existing methods and baselines, the effectiveness and efficiency of our method and its components are demonstrated. Sensitivity analysis and case studies are also included in this section. All the

experiments were conducted on a computer with one 3.20GHz Intel Xeon CPU and 18.0 GB RAM.

2.3.1 Dataset and Labels

Twitter data used in this work was purchased from Datasift Inc. (www.datasift.com). All analyses here are done in compliance with Twitter and Datasift terms of use. The dataset consists of randomly selected 10% tweets of all the tweets sent in the period from July 2012 to May 2013 in the 10 countries listed in Table 3.2. This dataset was separated into two parts: 1) data from July to October 2012, which served as the training set for the supervised comparison methods, and 2) data from November 2012 to May 2013, which was used as the testing set for validating all the methods. Both the training set and testing set were partitioned into date intervals and event detection was performed for each country individually based on each day's data. Stop-words from tweets were eliminated while stemming was also implemented.

Table 2.3: Dataset and Label Source

Country	**#Tweets (million)**	**News source** [1]	**#Events**
Argentina	52	Clarn; La Nacin; Infobae	365
Brazil	57	O Globo; O Estado de So Paulo; Jornal do Brasil	451
Chile	28	La Tercera; Las ltimas Notcias; El Mercurio	252
Colombia	41	El Espectador; El Tiempo; El Colombiano	298
Ecuador	13	El Universo; El Comercio; Hoy	275
El Salvador	7	El Diro de Hoy; La Prensa Grfica; El Mundo	180
Mexico	51	La Jornada; Reforma; Milenio	1217
Paraguay	8	ABC Color; Ultima Hora; La Nacon	563
Uruguay	3	El Pa; El Observador	124
Venezuela	45	El Universal; El Nacional; Ultimas Notcias	678

Our detection results were validated against a labeled events set, namely the Gold Standard Report (GSR). GSR was exclusively provided by MITRE [74]. The general collection protocol followed by the GSR is as follows: for each country, the top 3 newspapers were selected from among the top 100 newspapers published in Latin America, as provided by International Media and Newspapers. News was also collected from the most influential international news outlets and with additional input from subject matter experts. An event was considered "significant" if it was reported by any of these news outlets. The dataset and labeled news sources for each of these countries are listed in Table 3.2.

[1] In addition to the top 3 domestic news outlets, the following news outlets are included: The New York Times; The Guardian; The Wall Street Journal; The Washington Post; The International Herald Tribune; The Times of London; Infolatam.

Table 2.4: Methods and Efficiencies

Methods	**Targeted Domain**	**Supervised**	**Running Time**
Earthquake Detection [90]	Yes	Yes	15.2 hours
Topic Modeling [106]	No	No	9.7 hours
Graph Partition [100]	No	No	18.9 hours
ST Burst [63]	No	No	30.1 hours
TEDAS [64]	Yes	Yes	20.9 hours
QE [71] +LMSS	No	No	23.2 hours
SVM+LMSS	Yes	Yes	22.0 hours
DQE+SS [77]	Yes	No	16.3 hours
Our proposed (DQE+LMSS)	Yes	No	18.2 hours
EDSS [4]	No	No	19.8 hours

2.3.2 Methods for Comparison

Table 2.4 lists all the comparison methods tested: Earthquake Detection, Topic Modeling, Graph Partition, Spatial Temporal Burst (ST Burst), TEDAS, and EDSS. Their implementations and parameters settings were as follows.

Earthquake Detection [90]: This method is initially proposed to detect earthquake, here it is borrowed to detect civil unrest events. 5,386 tweets were manually labeled as "civil unrest related" and another 6,147 tweets as non-related for training purposes. Three types of features were evaluated: statistical, keyword, and word context. All these types of features were tested and the keyword feature were chosen for its best performance.

Topic Modeling [106]: The implementation was provided by the authors. Hashtags were treated as tags and tweet geotags were deemed to be the corresponding geographic regions.

Graph Partition [100]: The authors employed a weighted Median Absolute Deviation to handle the skewness of the signal strength distribution. Various weight values from 1 to 40 were evaluated and the value 20 was chosen since it achieved the best performance.

ST Burst [63]: The implementation was provided by the authors. The tunable temporal window size was set to 6 in the original work. Other values are also evaluated, including 12 and 24, but observed similar results.

TEDAS [64]: The tunable parameters (α, β) and (α', β') were used to denote the priors to punish words with low frequencies. The well-recognized setting: $\beta = \beta' = 10$ was followed to filter out trivial words. The setting $\alpha = \alpha' = 0.1$ was adopted due to the low percentage of civil unrest content.

EDSS [4]: The tunable parameter, λ, balances the relative significance of content and network structure in event detection. Various values, including the extreme values of 0 and 1 for λ, were tested. The setting $\lambda = 0.5$ was adopted since it outperformed the other settings tested.

In addition, the effectiveness of each component of our DQE+LMSS was tested by comparing with those of 3 baselines, namely query expansion (QE)+LMSS, support vector machine

(SVM) +LMSS, and DQE+spatial scan (SS):

QE+LMSS: QE was implemented by following the original design in [71], and adopted the same seed query as that used in DQE.

SVM+LMSS: SVM was adopted by following the experimental settings of the "Earthquake Detection" method. Domain-related tweets were extracted based on SVM and then utilized by LMSS.

DQE+SS: As a popular spatial scan statistic, the Kulldorff statistic was applied with our DQE to compose a baseline [77].

2.3.3 Validation

All the comparison methods and baselines returned the event-related tweet content and the corresponding time and location. In addition to the "civil unrest" domain, the general-interest event detection methods output the events under any domains. Therefore, to achieve a fair comparison, events from other domains were filtered out for these methods. In particular, given the "event"-related tweets generated for any method, a linear SVM classifier was adopted to classify all the events into two categories: events in the civil unrest domain and those in other domains. The classifier utilized unigram features, and was trained based on 5,386 tweets manually labeled as "civil unrest related" and another 6,147 tweets labeled as "unrelated".

After the extraction of "civil unrest" events by the general-interest event detection methods, all the methods were validated against the GSR. A detected event "matches" a GSR event if the following conditions are both satisfied: 1) the event time detected is the same as the time period recorded in GSR; 2) the event location detected is within the same city as that recorded in GSR.

2.3.4 Initial Settings

There are several tunable parameters in our approach. β in Equation 2.2 is a parameter for updating tweet node weights. Its default value is set to 1. λ in Equation 2.7 was used to balance the weights between local modularity and spatial scan statistics, and with 1 as its default value. Other settings of β and λ were also studied and are discussed in the rest of the paper.

To initialize DQE, a user is asked to choose 10 civil unrest tweets. In those tweets, terms are ranked based on their document frequency-inverse document frequency (DFIDF) [100] weights. For Spanish speaking countries, the top keywords are: "protesta", "marcha", "movimiento", "patritica", "manifiesto", "violencia", "holgun", "americateve", "cubanet", and "rolezeiros" in a descending order of DFIDF. Based on our experiments, the top ranked

terms are generally related to civil unrest, such as "protesta", "marcha", and "movimiento", whatever the initial 10 civil unrest tweets selected. The same situation applies for Portuguese speaking countries. The top 5 keywords were selected as the seed query terms, all of which were assigned with the same weight to form the seed query. The impact of the number N of seed query terms is discussed in the *Study of Parameters* Section.

Additionally, in LMSS, the longest distance r between any two neighboring locations was set to 200km. 20 additional values of r were also tried, ranging from 150km to 370km, which was found to make little difference to the performance, as noted in the *Study of Parameters* Section.

2.3.5 Evaluation of Components

First, the empirical cases will be presented to illustrate the correctness of the expanded query generated by DQE, then the effectiveness of DQE and LMSS are demonstrated based on quantitative comparisons with the baseline methods.

Table 2.5: Performance Comparison with Baseline Components (Precision, Recall, F-measure)

Dataset	DQE+LMSS	DQE+SS	QE+LMSS	SVM+LMSS
Brazil	**0.93**, 0.37, 0.53	0.84, **0.59, 0.69**	0.44, 0.14, 0.21	0.39, 0.24, 0.30
Colombia	**0.81, 0.75, 0.78**	0.58, 0.73, 0.65	0.31, 0.16, 0.21	0.63, 0.64, 0.63
Uruguay	0.66, **0.82, 0.73**	0.76, 0.26, 0.39	**0.80**, 0.58, 0.67	0.45, 0.27, 0.34
El Salvador	**0.83, 0.43, 0.56**	0.63, 0.09, 0.16	0.55, 0.37, 0.44	0.61, 0.19, 0.29
Mexico	**0.91, 0.49, 0.6**4	0.73, 0.37, 0.49	0.56, 0.09, 0.16	0.56, 0.18, 0.27
Chile	**0.80**, 0.69, **0.74**	0.58, **0.75**, 0.65	0.28, 0.28, 0.28	0.78, 0.29, 0.42
Paraguay	**0.98**, 0.35, 0.52	0.96, 0.17, 0.29	0.88, **0.67**, **0.76**	0.57, 0.11, 0.19
Argentina	0.78, 0.61, 0.69	0.69, **0.71**, **0.70**	0.67, 0.54, 0.60	**0.92**, 0.22, 0.35
Venezuela	**0.88, 0.50, 0.64**	0.57, 0.31, 0.40	0.56, 0.26, 0.36	0.65, 0.12, 0.20
Ecuador	**0.82**, 0.51, 0.63	0.72, 0.44, 0.55	0.54, **0.93**, **0.68**	0.62, 0.71, 0.66

Quality Analysis of DQE's Performance

Here, DQE is proposed to generate the expanded queries. Table 2.6 lists GSR events in July 2012 in Mexico and the corresponding expanded query terms generated by DQE. In the second column, for each date the 6 query terms with the highest weights are listed as the representatives of each expanded query. For each date, the expanded query terms are not only all related to the civil unrest domain, but are also very relevant to the GSR description on that date. Determinative key terms such as event locations, event times, and organization names are successfully identified. Moreover, event-related key hashtags (e.g.,

Table 2.6: Comparison between Expanded Query from DQE and GSR Description of Events

Detect-Date	**Expanded Query Extracted**	**GSR Description of Real Events**	**Occur-Date**
1-Jul	#YoSoy132, #Granmarcha132, patrull, Companer, PRI, movement	"Youth movement #YoSoy132 staged a sit-in outside the local board of Federal Electoral Institute."	1-Jul
3-Jul	#Epnnuncaseramipresidente, fraud, #YoSoy132, movimient, progress, contig, march	"The student movement #YoSoy132 protested against fraud in the elections."	3-Jul
7-Jul	#Megamarcha, #Exigimosdemocracia Eugenio, Derbez, eleccion, @YoSoy132Media	"Protesters unite to call for mega march." "YoSoy132 go and concentrate on the Esplanade of Heroes."	7-Jul
8-Jul	#Megamarcha, #Megamarch, Eugenio, Derbez, against, election	"Protesters unite to call for mega march against virtual presidential election."	
13-Jul	imposicion, #Megamarcha, 15hrs, principal, march, #AMLO	"A march was in protest of the imposition of the PRI candidate."	14-Jul
14-Jul	#Megamarcha, #Megamarch, 14juli, zocal, angel, march	"Virtual #Megamarch against the winner of the presidential election, Enrique Pea Nieto, left the Angel de Independencia to el Zocalo of Mexico City."	14-Jul
19-Jul	#Sosmexico, #Sosmexic, fraud, elector, march, protest	"Protesting for alleged fraud in the election of July 1"	19-Jul
22-Jul	#Megamarcha, #YoSoy132, @epigmenioibarra, Zocal, march, imposicion	"A mega march against the alleged imposition of the PRI." "YoSoy132 march arrives at El Zocalo and goes to the Monument to the Revolution"	22-Jul
27-Jul	#Ocupatelevisa, #YoSoy132, televisa, chapultepec, installation, march	"Students symbolically take over facilities of Hidalgo Radio and TV, and fence outside Televisa Chapultepec in Mexico City"	27-Jul

Table 2.7: Performance Comparison with Existing Event Detection Methods (Precision, Recall, F-measure)

Dataset	DQE+LMSS	Graph Partition	Earthquake	Topic Modeling	TEDAS	ST Burst	EDSS
Brazil	**0.93**, 0.37, 0.53	0.55, 0.34, 0.42	0.65, 0.19, 0.30	0.46, 0.09, 0.15	0.39, 0.20, 0.27	0.80, **0.45**, **0.58**	0.86, 0.28, 0.42
Colombia	0.81, **0.75, 0.78**	0.68, 0.29, 0.41	0.55, 0.49, 0.52	0.26, 0.39, 0.31	0.66, 0.41, 0.50	**0.87**, 0.48, 0.62	0.57, 0.52, 0.54
Uruguay	0.66, **0.82, 0.73**	0.28, 0.23, 0.25	0.86, 0.11, 0.20	0.22, 0.06, 0.09	**0.88**, 0.56, 0.68	0.11, 0.06, 0.08	0.66, 0.13, 0.22
El Salvador	**0.83, 0.43, 0.56**	0.35, 0.07, 0.10	0.32, 0.06, 0.10	0.40, 0.05, 0.09	0.71, 0.36, 0.48	0.30, 0.12, 0.17	0.52, 0.15, 0.23
Mexico	**0.91, 0.49, 0.64**	0.72, 0.23, 0.35	0.51, 0.19, 0.28	0.34, 0.08, 0.12	0.56, 0.20, 0.29	0.76, 0.43, 0.55	0.69, 0.27, 0.39
Chile	0.80, **0.69, 0.74**	0.83, 0.39, 0.53	0.46, 0.19, 0.27	0.42, 0.48, 0.45	**0.96**, 0.36, 0.53	0.67, **0.69**, 0.68	0.35, 0.43, 0.39
Paraguay	**0.98**, 0.35, 0.52	0.76, 0.19, 0.30	0.40, 0.10, 0.16	0.86, 0.07, 0.13	0.88, **0.67, 0.76**	0.34, 0.12, 0.18	0.83, 0.16, 0.27
Argentina	0.78, 0.61, **0.69**	**0.88**, 0.14, 0.24	0.63, 0.57, 0.60	0.38, 0.42, 0.40	0.51, **0.64**, 0.57	0.63, 0.73, 0.67	0.73, 0.55, 0.63
Venezuela	**0.88, 0.50, 0.64**	0.46, 0.21, 0.29	0.87, 0.22, 0.35	0.47, 0.37, 0.41	0.79, 0.28, 0.42	0.82, 0.33, 0.47	0.86, **0.50**, 0.63
Ecuador	**0.82**, 0.51, 0.63	0.30, 0.22, 0.25	0.78, **0.60**, 0.68	0.67, 0.04, 0.08	0.55, 0.92, **0.69**	0.29, 0.26, 0.27	0.64, 0.28, 0.39

"#Megamarcha") and keyplayers (e.g., "epigmenioibarra") were also effectively extracted. Interestingly, the only exception was on July 8th, where the key term "Eugenio Derbez", a popular celebrity in Mexico, was detected. This name became a key term because the protest happened to occur near to his wedding venue, which was reported in online media.

Quantitative Analysis of DQE's Effectiveness

The experiments focus on examining whether DQE is the best choice for our event detection method, compared to other classic methods. Therefore two baseline options, QE and SVM, were introduced as potential replacements for DQE to be used in conjunction with LMSS. The performance of these baselines was then compared with our proposed DQE+LMSS. The results are shown in Table 2.5. DQE+LMSS achieved the best F-measures in 8 of the 10 countries and was second best in Paraguay and Ecuador. Moreover, it consistently achieved highly competitive F-measures of above 0.5 across all the countries tested, which confirms the stability of its performance. This demonstrates that DQE is a better choice for our event detection method.

Effectiveness of LMSS

In this set of experiments, the effect of utilizing LMSS as a component of our method is evaluated by comparing its performance against that of the baseline method DQE+SS described above. The results of the comparison are shown in Table 2.5. DQE+LMSS clearly outperforms DQE+SS, achieving much higher F-measures in most of the countries tested except Brazil and Argentina. DQE+SS has F-measure values below 0.5 in half of the countries, and its recall values are lower than 0.3 in 3 countries. The superior performance demonstrated by both DQE and LMSS vindicate the decision to utilize them as the components of the proposed event detection method.

2.3.6 Event Detection Performance

Our proposed approach was compared with existing methods based on precision, recall, and F-measures on civil unrest event detection.

The experimental results are illustrated in Table 2.7, which shows that the proposed method achieves the best overall performance. Except for Brazil, Ecuador, and Paraguay, DQE+LMSS achieves the highest F-measures in every country. Even for these 3 countries, it scored the best on precision and achieved a highly competitive overall performance. Although TEDAS also achieves a relatively good performance compared to the other benchmark methods, it still produced 4 countries with F-measures lower than 0.5. Among the existing methods, the Earthquake method and EDSS were relatively advantageous in precision, but suffered from a limited ability to detect most of the events. ST Burst performed better in large countries such as Brazil, Argentina, and Mexico, than in the smaller ones. Graph Partition and Topic Modeling, which are unsupervised methods designed for events under general-interest domain, seem relatively weak for detecting events under a targeted domain, achieving F-measures over 0.5 in very few countries.

The computation times consumed by these methods are shown in Table 2.4. There is no significant difference in running times among most of the methods. The only exception is Topic Modeling, which took less than 10 hours. Note that unlike the other targeted-domain spatial event detection methods, namely Earthquake and TEDAS, our method is unsupervised, which means it does not need to devote additional effort to labeling.

In summary, the experiments clearly demonstrate the effectiveness and efficiency of the proposed DQE+LMSS approach.

2.3.7 Study of Parameters

The impact of the parameters of the proposed approach was evaluated, including (i) N, the number of the seed query terms, (ii) β, the parameter for updating tweet node weights (see Equation 2.2), (iii) λ, the trade-off between local modularity and spatial scan statistics (see Equation 2.7), and (iv) r, the longest distance between any two neighbor nodes.

Figure 2.3(a) illustrates the performance of our method versus N, the number of seed query terms. For most of the countries, the F-measures corresponding to $N = 2$ or 3 are significantly higher than when $N = 1$. But when N increases further, the F-measures tend to be stable, especially once N reaches 5.

Figure 2.3(b) shows the results of varying β from 0.5 to 1.5. By increasing the value of β to around 1, the F-measures of most countries are improved, but once it exceeds 1, the performance drops.

The results of tuning λ are shown in Figure 3(c). By varying λ from 0.3 to 1.4, the F-

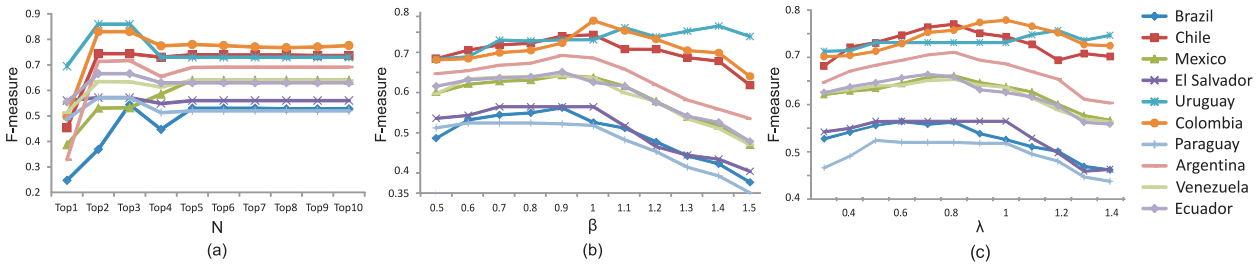


Figure 2.3: Sensitivity analysis of parameters. (a) Sensitivity analysis of "number of seed query terms" (b) Sensitivity analysis of "trade-off β for updating tweet node weights" (c) Sensitivity analysis of "trade-off λ between local modularity and spatial scan statistics".

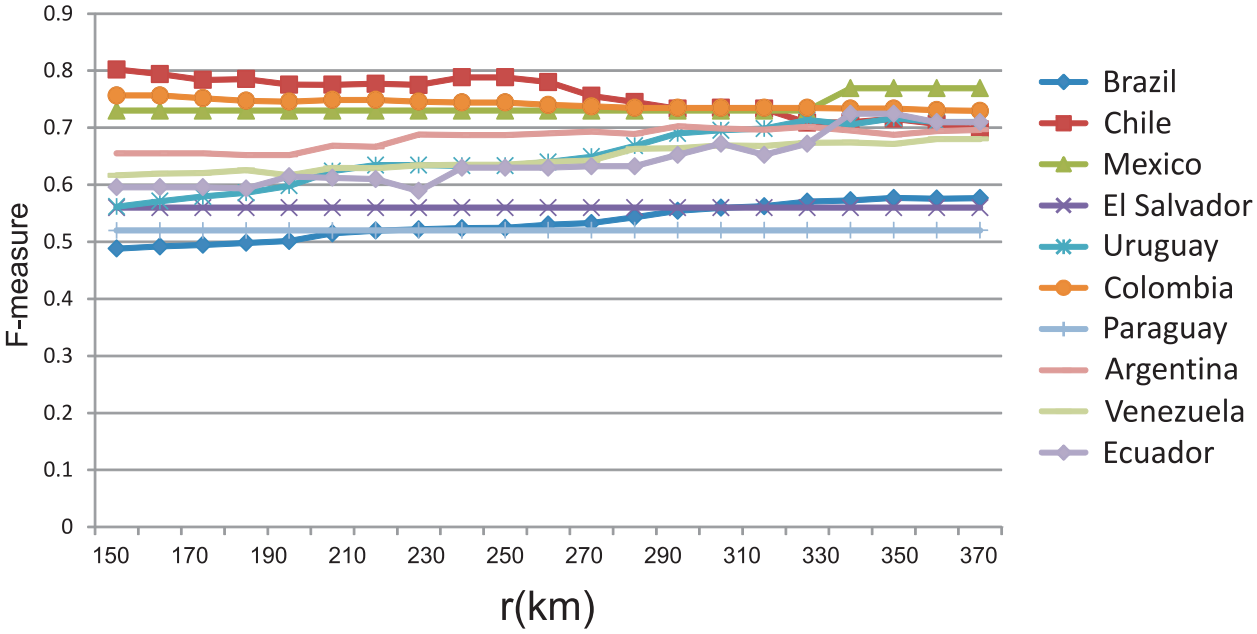


Figure 2.4: Sensitivity analysis of the longest distance r between any two neighboring locations.

measures generally increase, reaching their peaks when λ is in the range of 0.7 to 1.2. This suggests the "sweet region" of λ to correspond to the point where the local modularity and spatial scan statistics combine to achieve the optimal performance. Moreover, even with an extreme value of λ, say 0.3 or 1.4, the overall performance of the proposed model remains highly competitive compared to its peers, as can be seen in the data shown in Table 2.7.

Figure 2.4 illustrates the F-measures obtained by varying r from 150km to 370km. The F-measures for Colombia, Paraguay, Mexico, El Salvador, and Argentina do not change significantly with respect to r. For the other 5 countries, the effect on the F-measures of varying the value of r are mostly less than 0.1, which are still minimal.

2.3.8 Case Study

During the experiments, a number of interesting facts revealed by using the proposed approach was observed. For instance, comparing the results for Colombia and Paraguay, the very different expanded query terms from these two countries reflect their correspondingly different social foci, which contributed to the model's ability to accurately detect local events accordingly. As shown in Figure 2.5, the major term for movements is "protest" in Colombia (as on October 1, 2012) versus "huelga" (i.e., "strike" in English) in Paraguay (as on November 20, 2012). The cities "Medellin" in Colombia and "Curuguaty" in Paraguay were both hot spots for unrest events, but the movements in Colombia seem more metropolitan-related, because of the appearance of terms such as "estacion" (station), "transport", and "teleantioqui" (television). Paraguay's themes for these events are more about "libert", "campesin" (peasant), and "hambr" (hunger). These cases reveal that our method can indeed capture the variety of keywords across different countries. It is worth noting that ongoing unrest keywords, even in the same country, tend to evolve over time, as shown in Table 2.6, and our DQE can still capture this evolution effectively.

Based on the expanded queries generated by our DQE, LMSS was able to identify spatial unrest events. In the above examples, as shown in Figure 2.5, the proposed method detected one event on October 1st, 2012 in Colombia that was related to transportation in the city of Medellin; on Nov. 20th, 2012 in Paraguay, the proposed method detected 2 events concerned with about "food subsidy" in Curuguaty and "peasants demand freedom" in Asunci n, respectively.

2.4 Conclusion

This chapter presents a novel unsupervised approach for detecting spatial events under targeted domains. dynamic query expansion is developed that utilizes a Twitter heterogenous information network to dynamically extract domain-related key terms. To extract spatial events based on these domain-related tweets, we designed a local modularity spatial scan capable of simultaneously considering the semantic similarity and the geographical proximities of tweets. Extensive empirical studies on civil unrest event detection were conducted based on Twitter data collected in 10 Latin American countries. The results demonstrated the effectiveness and efficiency of our proposed approach.

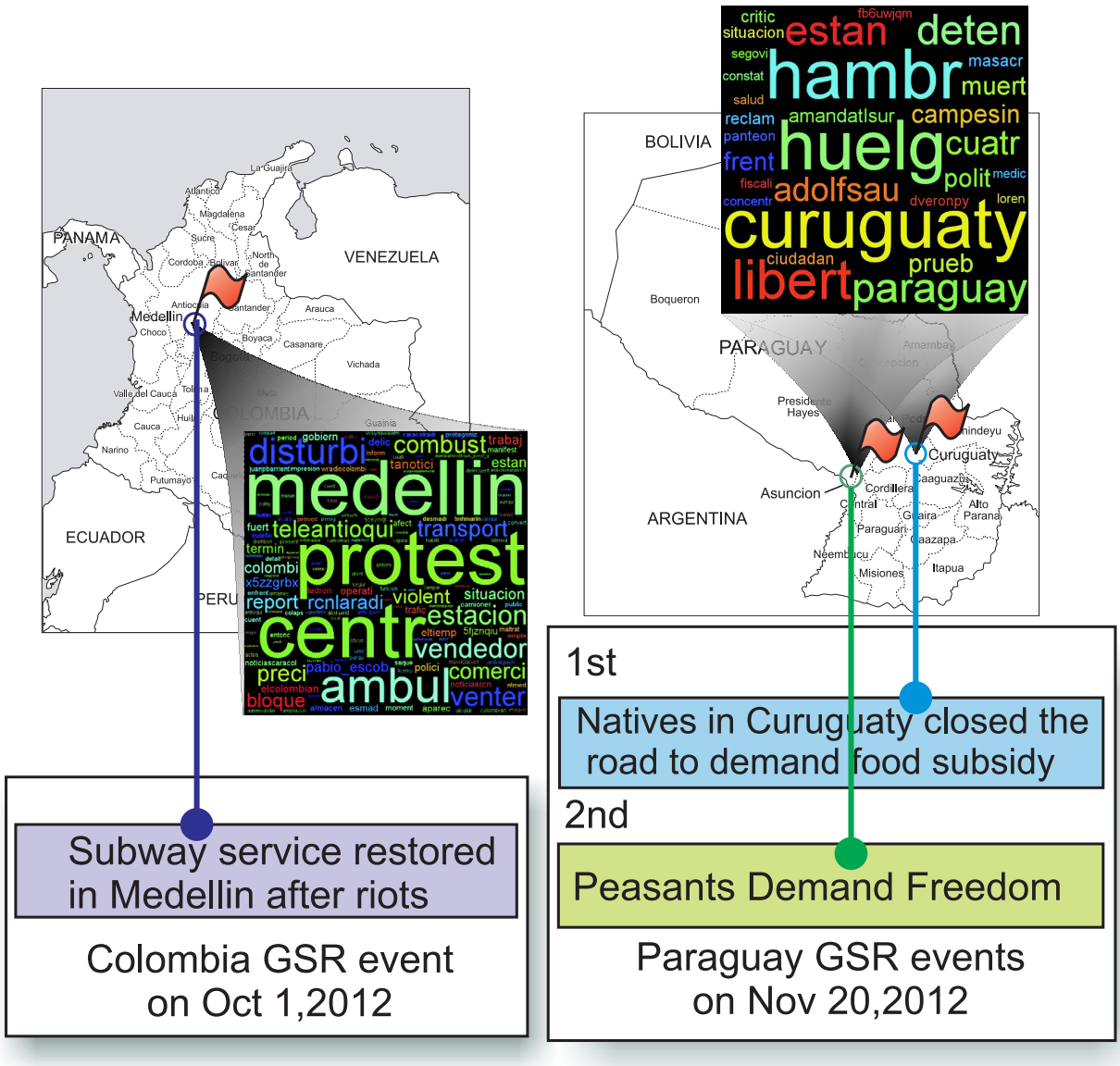


Figure 2.5: Event detection case studies.

Chapter 3

A Generative Framework for Event Forecasting

In this chapter, I propose a generative model for spatiotemporal event forecasting in Twitter. The model characterizes the underlying development of future events by jointly modeling the structural contexts and spatiotemporal burstiness. Section 3.1 presents the introduction of the research of this chapter. Section 3.2 reviews existing work. Section 3.3 describes the proposed generative model and associated parameter estimation details. Section 3.5 explains the event forecasting function of the proposed model. In Section 3.6, extensive experiments to evaluate the performance of the new model are conducted and analyzed; the work is summarized and conclusions drawn in Section 3.7.

3.1 Introduction

Microblogs like Twitter and Weibo are important platforms to reveal and discuss social events [61]. As of the end of 2014, 500 millions of tweets are posted everyday mainly by 255 million active users, discussing a variety of content ranging from everyday feelings to comments about social events [15]. Compared to traditional media, Twitter has the following significant characteristics: 1) *Timeliness of messages:* Unlike traditional media that take hours or days to publish, tweets can be posted instantly utilizing portable mobile devices; 2) *Ubiquity of social sensors:* Tweets reflect the public's mood and trends, which could be the determinants of future social events; and 3) *Availability of geo-information:* Twitter users provide rich location information in profiles, texts, and geotags. Recent research has revealed the power of Twitter for event forecasting [96,99]; Twitter and other social media have been recognized for playing a key role in events such as the "Arab Spring" and the Mexican presidential election protests [86, 99]. Figure 3.1 depicts activities on Twitter that causally preceded the Mexico City protests. Both the content and spatiotemporal burstiness of the protest-

related tweets reveal the escalation of societal discontent pertaining to this controversial election, from complaining through planning and advertising, to the final protest event. However, existing event forecasting models in Twitter generally focus on temporal events whose geo-locations are not available or irrelevant to the prediction task (e.g., elections [96] and sports [84]). Comparatively little attention has been paid to forecasting spatiotemporal events.

A spatiotemporal event is mainly relevant to the tweets posted within a certain geographical neighborhood. Thus, the forecasting of spatiotemporal events requires a consideration of spatial features and their correlations in addition to the temporal dimension. This poses the following three challenges: 1) *Capturing spatiotemporal dependencies.* A spatial event may influence not only the location and time, but also its geographical and temporal neighborhood. The influence strength and pattern may vary in different development stages for different events; 2) *Modeling mixed type observations.* An event involves the temporal evolution of spatially distributed tweets and their semantics. Joint consideration of these heterogeneous and multi-dimensional data is crucial; and 3) *Utilizing prior geographical knowledge.* Spatiotemporal events in crucial domains usually have rich historical records. Different geolocations may feature their inherent and distinct event frequencies that can be integrated into a predictive model to improve its forecasting accuracy. For example, the historical crime rates in different cities can help forecast the probability of future crime events.

This work proposes spatiotemporal event forecasting models that effectively address the above-mentioned issues. The proposed model generatively characterizes the evolutionary development of events, as well as the relationships between the tweet observations inside and outside the event venue. To uncover the underlying event development mechanics, the model jointly considers the structural semantics and spatial-temporal burstiness patterns in Twitter streams. Utilizing the geographical prior allows spatial burstiness distributions to be learned for corresponding locations. Applying a Gaussian-inverse Wishart prior distribution facilitates event forecasting for unknown locations. The main contributions of this work are:

- **A novel generative framework for spatial event forecasting.** For spatial event forecasting in Twitter, this chapter proposes an enhanced hidden Markov model (HMM) that characterizes the transitional process of event development by jointly considering the time-evolving context and space-time burstiness of Twitter streams.
- **Effective batch and online algorithms for model parameter inference.** The model inference is formalized as the maximization of a posterior that is analytically tractable. Both EM-based algorithm and stochastic-EM parameter optimization algorithms are proposed to solve this problem effectively and efficiently.
- **A new sequence likelihood calculation method.** To handle the noisy nature of tweet content, words exclusive to a single event are identified by a language model that is optimized by a dynamic programming algorithm to achieve accurate sequence likelihood calculation.

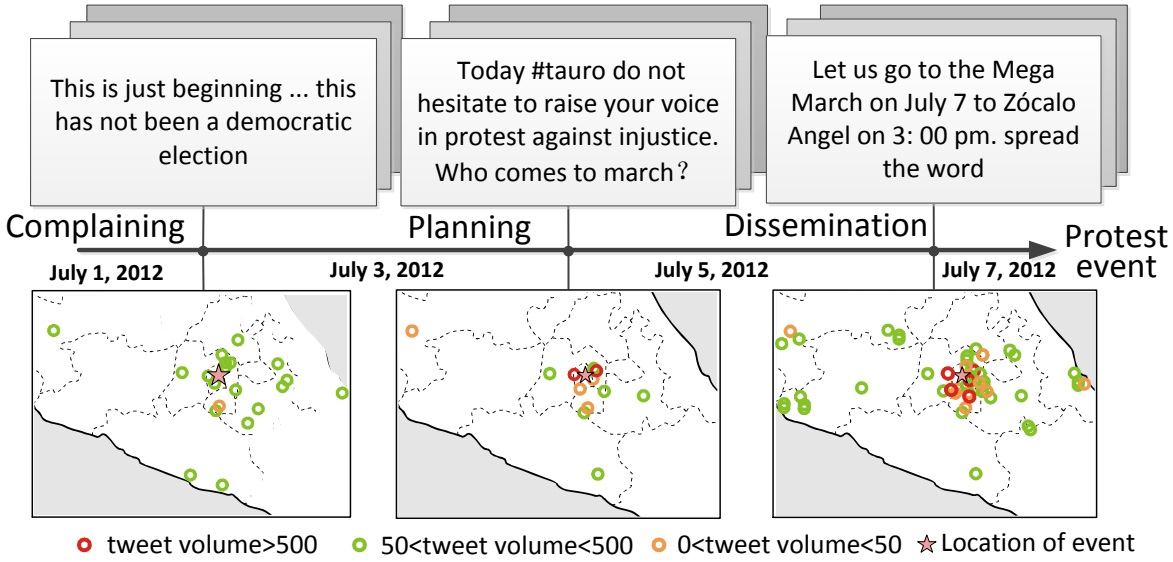


Figure 3.1: Twitter predicts a presidential election protest.

- **An online parameter optimization algorithm.** In addition to traditional , this work also develops online parameter optimization algorithm. It not only reduces the training expense on large dataset but also improves the timeliness of the model in the social media streams.

- **Extensive experimental performance evaluations.** The proposed method outperforms existing methods by 38% and 67% on two different datasets. Sensitivity analyses reveals the impact of the parameters on the new method's performance. Case studies on both datasets are illustrated and elaborated to demonstrate the practical usefulness of the proposed methods.

3.2 Related Work

Current researches into the analysis of Twitter-based social events can be categorized into two main types: 1) event detection; and 2) event forecasting. These are considered in turn below.

Event detection: A large body of work focuses on the detection of ongoing events [4,63,90, 92,100]. They utilize tweets as real-time and ubiquitous social sensors to promptly discover new events occurring. Methods based on spatial bursts use a classifier to extract topic-related tweets and then examine their spatial burstiness, in applications such as detecting earthquakes [90] and disease outbreaks [92]. Methods based on temporal bursts detect the

temporal patterns of Twitter streams utilizing techniques such as wavelet analysis [100] or temporal clustering [4]. Spatiotemporal methods aim to detect bursts in both time and space [63,109]. However, these event detection approaches can only uncover events after they have occurred and are unable to forecast future events because they all focus on observations that directly reflect currently occurring events, rather than precursor indicators that reveal the causes or development of future events.

Event forecasting: Most research in this area focuses on temporal events and ignores the underlying geographical information. A variety of applications have been explored, including elections [78, 96], disease outbreaks [2, 89], stock market movements [9, 21], politics [70], box office ticket sales [9], the Olympic games [84], crime [99], and traffic conditions [49]. These papers can be categorized into four types based on the complexity of models utilized: **1)Linear regression model.** This thread maps simple predictive features such as sentiment score or tweet volume to the occurrence of future events [9,21,49,78]; **2) Nonlinear models.** This thread incorporates more informative features such as semantic topics by utilizing methods such as support vector machines and logistic regression [89, 99]; **3) Time series-based methods.** This thread considers the temporal correlation of relevant features such as tweet volume by adopting approaches such as autoregressive modeling [2]; and **4) Domain-specific approaches.** This thread is designed to solve particular problems and may not be applicable to other application domains. For example, Pavlysehko [84] applied an association rule approach to discover the most frequently mentioned players and hence predict the results of sports tournaments, while Marchetti-Bowick and Chambers [70] focused on improving the performance of sentiment analysis related to political events. As yet, there have been few reports of work specifically on spatiotemporal event forecasting. Gerber [41] proposed a predictor for spatiotemporal events by utilizing historical event counts and topics, but do not consider temporal evolution and dependencies, while Wang et al. [98] developed a model to characterize and predict spatio-temporal criminal incidents, but their model requires the availability of demographic information. Zhao et al. [112] proposed three multitask learning models to forecast civil unrest events utilizing static features and dynamic features. Instead of considering geographic neighborhood, it assumes all the locations in a country are equally interactive to each other.

This work proposes a spatiotemporal event forecasting method that characterizes the evolutionary pattern of both spatial burstiness and structural contexts. By modeling geographical priors effectively, the new approach can sufficiently leverage historical prior knowledge and can be effectively applied to new locations.

3.3 Generative Process of the Proposed Models

This section elaborates the formulation and generative process of the proposed methods. First, the spatiotemporal event forecasting problem is formalized. Then, the proposed new generative model is described in detail, including the space-time burstiness module and

Table 3.1: Notations and descriptions

Notations	Descriptions
$Z_{s,t}$	Latent state in sequence s at time t.
$Y_{s,t,n}$	Category-switching variable of the nth word in sequence s at time t.
$X_{s,t,n}$	Topic label of the nth word at time t in sequence s.
$W_{s,t,n}$	The nth word in sequence s at time t.
$r_{s,t}^{in}$	The posting ratio in sequence s's location at time t.
$r_{s,t}^{out}$	The posting ratio outside the location of sequence s at time t.
$N_{s,t,w}$	The frequency of a word w in sequence s at time t.
Ψ	Bernoulli distribution that generates $Y_{s,t,n}$.
Φ	Topic distribution that generates $X_{s,t,n}$.
θ_j^B	Distribution of words under the jth topic.
$\theta s, t^R$	Distribution of words exclusive to sequence s at time step t.
$\mu_{l,k}$	Mean of posting ratios of location l under latent state k.
$\Sigma_{l,k}$	Covariance of posting ratios of location l under latent state k.
$\lambda_{l,k}^{in}$	Mean of posting ratios inside the location l for Poisson distribution for latent state k.
$\lambda_{l,k}^{out}$	Mean of posting ratios outside the location l for Poisson distribution for latent state k.

structural tweet content module.

3.3.1 Problem Formulation

The notations used in the paper are introduced in Table 3.1. As demonstrated in Figure 3.1, to accurately forecast spatiotemporal events it is crucial to be able to characterize their underlying development before the occurrence by utilizing relevant tweet observations. An enhanced hidden Markov model is proposed here to characterize the underlying development of events.

Given a sequence of observations (i.e. symbols) O, a standard HMM can be denoted as a quadruple (H, Z, A, π), where Z is a set of K latent states. $H_k(O_i)$ denotes the emission probability that a symbol O_i is generated by the kth latent state. A is a $K \times K$ transition probability matrix, where $A_{j,k} = p(Z_j|Z_k)$is the transitional probability of moving from the jth latent state to the kth latent state and π is the initial probability vector where π_k is the probability that the initial state is k. Starting from an initial state k, the HMM generates an observation O_1 according to the emission probability $H_k(O_1)$, and then transitions to a state j with the transitional probability $A_{j,k}$. The training process for an HMM thus entails searching for the set of parameters (H, Z, A, π) that best fit the sequence of observations.

However, a standard HMM is limited to simple symbol observations and will thus face several

challenges in our case as the observation does not consist of a single symbol but rather all the domain-related tweets in each time step. Further, a standard HMM can neither characterize spatial burstiness nor handle structural and noisy observations. Here, both the content and the spatial burstiness of domain-related tweets are the observations, and the underlying stage in the development of social events is characterized as the latent state. A future event is predicted by inferring the underlying development with tweet observations.

This problem therefore requires several important enhancements to the standard HMM. First, instead of a single symbol, each observation encompasses all the domain-related tweets in each time step. Second, the enhanced HMM treats the spatial burstiness of domain-related tweets as multivariate "posting rates" in the same geographical neighborhood. Third, to address the noisy nature of tweet content, a language model is used to filter out typos and identify proper names exclusive to particular events. Fourth, the structural semantics of the filtered tweets is modeled as a mixture of latent topics. The generative process of the new model is described in the following subsections.

More formally, denote $D = \{D_{l,t}\}_{l\in\mathcal{L},t\in\mathcal{T}}$ as a collection of space-time-indexed Twitter data split into different geographical locations $\mathcal{L}$ and different time intervals $\mathcal{T}$. A sequence of tweets is defined as $s = \{D_{l,t}\}_{t\in T\subseteq\mathcal{T}}$, which contains all the tweets in location l in the time period $T \subseteq \mathcal{T}$. S denotes the number of all such sequences in the data D. The proposed new model characterizes the development of each event as a sequence of latent states $Z = \{1, 2, \cdots, K\}$, with tweet sequence $s \subseteq D_l$ being the observations generated by the latent states.

Structural tweet content modeling

In domain-related tweet content, a word is deemed to belong to one of two categories: 1) Specific words: These are specific to a unique event, such as hashtags, hyperlinks, landmarks, and organization names; 2) Common words: These words are commonly used by different events, especially those that reflect the stage of development. In the kth latent state, the probability that a word belongs to either of the above two types is modeled by a Bernoulli distribution:

$$Y_{s,t,n} \sim Bern(Y_{s,t,n}|\Psi_k) \tag{3.1}$$

If a word $W_{s,t,n}$ in sequence s at time step t belongs to the first category, it is directly generated from a language model $\theta^R_{s,t}$, which designates the words exclusive to the current observation sequence s at current time t:

$$W_{s,t,n} \sim Mult(W_{s,t,n}|\theta^R_{s,t}) \tag{3.2}$$

If the word belongs to the second category, then it is selected from one of the latent topics that are shared by all such events.

$$X_{s,t,n} \sim Mult(X_{s,t,n}|\Phi_k) \tag{3.3}$$

A latent topic j is modeled as a multinomial distribution over words:

$$W_{s,t,n} \sim Mult(W_{s,t,n}|\theta^{B_j}) \tag{3.4}$$

3.3.2 Model I: Space-Time Burstiness Modeling with Neighborhood Interactions (STM-I).

The proposed model STM-I characterizes the space-time burstiness by not only handling the signal strengths on different locations, but also the potential correlations among these locations by leveraging the covariance.

Given a tweet sequence $s \subseteq D_l$ in location l, denote $c_{s,t}^{in}$ as the count of domain-related tweets inside location l at time t, and $c_{s,t}^{out}$ as the count outside this location; Denote $b_{s,t}^{in} = |D_{l,t}|$ as the total tweet count inside the location l at time step t, and $b_{s,t}^{out}$ as that outside this location. $r_{s,t}^{in} = c_{s,t}^{in}/b_{s,t}^{in}$ and $r_{s,t}^{out} = c_{s,t}^{out}/b_{s,t}^{out}$ are the *inside ratio* and the *outside ratio* and are, respectively, the proportions of the domain-related tweets inside and outside the location l. Hence, the spatial burstiness pattern surrounding the location l is jointly characterized by $r_{s,t}^{in}$ and $r_{s,t}^{out}$. For example, spatial burstiness typically occurs when the inside ratio is higher than the outside one. To characterize the spatial burstiness in terms of the inside and outside ratios, a bivariate Gaussian is utilized:

$$r_{s,t}^{in}, r_{s,t}^{out} \sim \mathcal{N}(r_{s,t}^{in}, r_{s,t}^{out} | \mu_{l,k}, \Sigma_{l,k}) \tag{3.5}$$

The advantages of a bivariate Gaussian are two fold. First, its covariance matrix quantifies the different significance of the inside and outside ratios in characterizing the spatial burstiness. Second, the non-diagonal elements of the covariance matrix can also capture the relationship between the inside and outside ratios.

For the kth latent state, draw the mean of the inside and outside ratios $\mu_{l,k}$ from a Gaussian distribution:

$$\mu_{l,k} \sim \mathcal{N}(\mu_{l,k} | \mu_0, \Sigma_{l,k}/\beta_0) \tag{3.6}$$

where μ_0 is the historical prior mean of the inside and outside ratios and β_0 is the number of prior measurements. $\Sigma_{l,k}$ is the scale matrix following the inverse Wishart distribution:

$$\Sigma_{l,k} \sim \mathcal{IW}(\Sigma_{l,k} | \Lambda_0^{-1}, \nu_0) \tag{3.7}$$

where Λ_0 and ν_0 describe the prior scale matrix and the degree of freedom, respectively.

As shown in Figure 3.2, the generative process of the proposed STM-I, which is the Gaussian-distributed burstiness modeling, is:

- For each sequence s at each time step t,
  - Draw $Z_{s,t} \sim Multi(Z_{s,t} | Z_{s,t-1}, A)$
- For each latent state k in each location l,
  - Draw the mean of the spatial burstiness from a normal distribution $\mu_{l,k} \sim \mathcal{N}(\mu_{l,k} | \mu_0, \Sigma_{l,k}/\beta_0)$

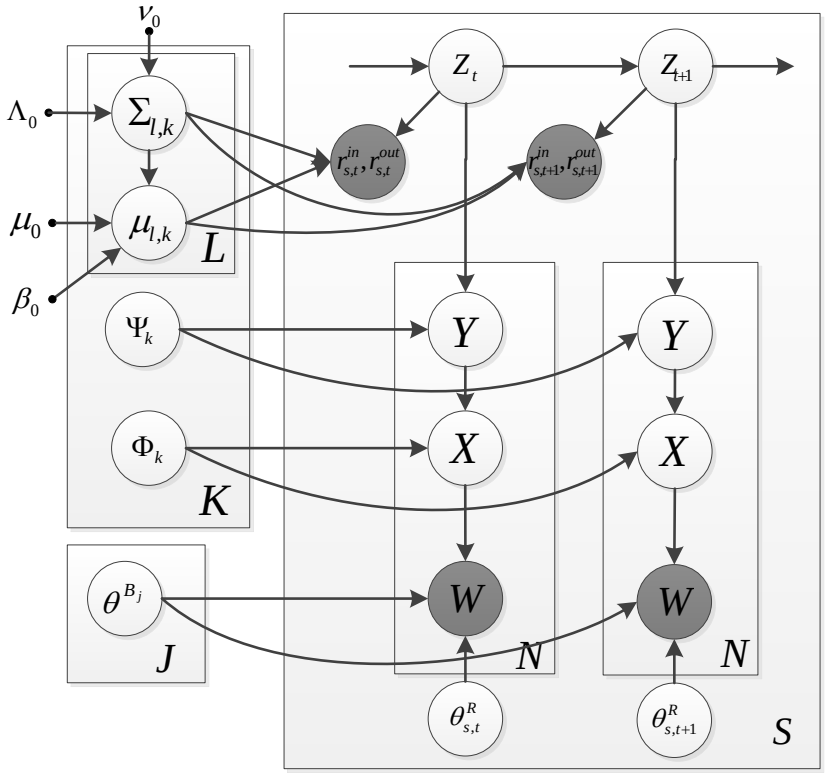


Figure 3.2: The plate notation of the proposed STM-I.

  - Draw the regional variance from an inverse Wishart distribution $\Sigma_{l,k} \sim \mathcal{IW}(\Sigma_{l,k}|\Lambda_0^{-1}, \mu_0)$
  - For each sequence of tweets s
    * Draw $r_{s,t}^{in}, r_{s,t}^{out} \sim \mathcal{N}(r_{s,t}^{in}, r_{s,t}^{out}|\mu_{l,k}, \Sigma_{l,k})$
- For each word W_n in time step t in tweet sequence s,
  - Draw $Y_{s,t,n} \sim Bern(Y_{s,t,n}|\Psi_k)$
  - If $Y_{s,t,n} = 0$, draw $W_{s,t,n} \sim Mult(W_{s,t,n}|\theta_{s,t}^R)$
  - else
    * Draw a topic $X_{s,t,n} \sim Mult(X_{s,t,n}|\Phi_k)$.
    * Draw a word $W_{s,t,n} \sim Mult(W_{s,t,n}|\theta^{B_j}, j = X_{s,t,n})$.

3.3.3 Model II: Space-Time Burstiness Modeling with Nonnegative-Discrete Signals.

Important signals such as the volume of tweets are typically nonnegative and discrete. The proposed STM-S is able to model the space-time burstiness by preserving these properties in the distribution assumptions.

Apart from Gaussian distribution, Poisson distribution is also commonly utilized in modeling the space-time burstiness [77]. The utilization of Poisson distribution will ensure the generated counts to be positive. Specifically, when assuming the count $c_{s,t}^{in}$ and $c_{s,t}^{out}$ are Poisson distributed with:

$$\begin{aligned} c_{s,t}^{in} &\sim Poisson(c_{s,t}^{in}|\lambda_{k,l}^{in} \cdot b_{s,t}^{in}), \\ c_{s,t}^{out} &\sim Poisson(c_{s,t}^{out}|\lambda_{k,l}^{out} \cdot b_{s,t}^{out}) \end{aligned} \tag{3.8}$$

where $b_{s,t}^{in}$ and $b_{s,t}^{out}$, the same as the above, represents the total tweet count inside and outside the location l at time step t, respectively. $\lambda_{k,l}^{in}$ and $\lambda_{k,l}^{out}$ denote the means of inside and outside outbreaks ratios, respectively.

The prior knowledge of the sufficient statistics of Poisson distribution for different locations and different states follow Gamma distributions:

$$\begin{aligned} \lambda_{k,l}^{in} &\sim Gamma(\lambda_{k,l}^{in}|\alpha^{in}, \beta^{in}), \\ \lambda_{k,l}^{out} &\sim Gamma(\lambda_{k,l}^{out}|\alpha^{out}, \beta^{out}) \end{aligned} \tag{3.9}$$

where α^{in} and β^{in} denote the shape parameter and inverse scale parameter of the Gamma prior for the inside outbreaks distribution. α^{out} and β^{out} denote the shape parameter and inverse scale parameter of the Gamma prior for the outside outbreaks distribution.

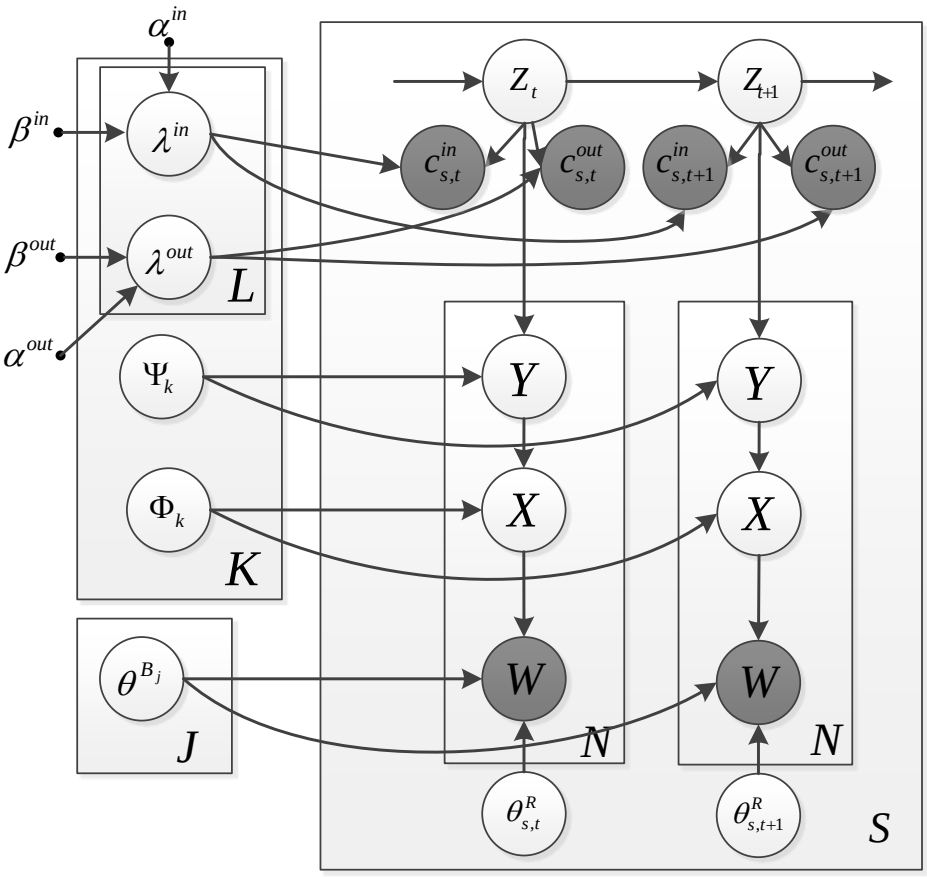


Figure 3.3: The plate notation of the proposed STM-S.

As shown in Figure 3.2, the generative process of the proposed STM-S, which is based on Poisson-distributed burstiness modeling, is:

- For each sequence s at each time step t,
  - Draw $Z_{s,t} \sim Multi(Z_{s,t}|Z_{s,t-1}, A)$
- For each latent state k in each location l,
  - Draw the mean of the in-location burstiness from a Gamma distribution $\lambda^{in}_{k,l} \sim Gamma(\lambda^{in}_{k,l}|\alpha^{in}, \beta^{in})$
  - Draw the mean of the out-location burstiness from a Gamma distribution $\lambda^{out}_{k,l} \sim Gamma(\lambda^{out}_{k,l}|\alpha^{out}, \beta^{out})$
  - For each sequence of tweets s
    - Draw $c^{in}_{s,t} \sim Poisson(c^{in}_{s,t}|\lambda^{in}_{k,l} \cdot b^{in}_{s,t})$
    - Draw $c^{out}_{s,t} \sim Poisson(c^{out}_{s,t}|\lambda^{out}_{k,l} \cdot b^{out}_{s,t})$
- For each word W_n in time step t in tweet sequence s,
  - Draw $Y_{s,t,n} \sim Bern(Y_{s,t,n}|\Psi_k)$
  - If $Y_{s,t,n} = 0$, draw $W_{s,t,n} \sim Mult(W_{s,t,n}|\theta^R_{s,t})$
  - else
    - Draw a topic $X_{s,t,n} \sim Mult(X_{s,t,n}|\Phi_k)$.
    - Draw a word $W_{s,t,n} \sim Mult(W_{s,t,n}|\theta^{B_j}, j = X_{s,t,n})$.

3.4 Parameter Estimation.

3.4.1 Joint Likelihood

Based on the generative process elaborated above, the proposed STM-I defines the joint probability of the generation of observed variables, latent variables, and model parameters.

Specifically, the observed variables are the spatial burstiness r^{in}, r^{out}, and words W in the tweet content; the latent variables are topic assignment X, category assignment Y, and latent state assignment Z. The geographical prior is $\Theta_I = \{\mu_0, \beta_0, \Lambda_0, \nu_0\}$. Their joint distribution

is expressed as follows:

$$
\begin{aligned}
&p(W, X, Y, Z, \mu, \Sigma, r^{in}, r^{out} | \pi, A, \Psi, \Phi, \theta, \Theta_I) \\
=&\prod_{s}^{S} p\left(Z_{s,1} | \pi\right) \cdot \prod_{s}^{S} \prod_{t=2}^{T} p\left(Z_{s,t} | Z_{s,t-1}, A\right) \\
&\cdot \prod_{s}^{S} \prod_{t=1}^{T} \prod_{n}^{N} p\left(W_{s,t,n}, Y_{s,t,n}, X_{s,t,n} | Z_{s,t}, \Psi, \Phi, \theta\right) \\
&\cdot \prod_{s}^{S} \prod_{t=1}^{T} p(r_{s,t}^{in}, r_{s,t}^{out} | \mu_l, \Sigma_l, Z_{s,t}) p(\mu_l, \Sigma_l | \Theta_I)
\end{aligned}
\tag{3.10}
$$

where $\theta = \{\theta^B, \theta^R\}$. Thus, searching for the best setting of the model parameters for STM-I is equivalent to the maximization of the logarithm of the joint distribution in Equation 3.10.

In STM-S, the Poisson-distributed space-time burstiness modeling is utilized. Specifically, the observed variables are the inside domain-related tweet counts c^{in}, the inside base counts b^{in}, the outside domain-related tweet counts c^{out}, the outside base counts b^{out}, and the words W in the tweet content; the latent variables are topic assignment X, category assignment Y, and latent state assignment Z. The geographical prior is $\Theta_{II} = \{\alpha^{in}, \beta^{in}, \alpha^{out}, \beta^{out}\}$. The joint distribution is expressed as follows:

$$
\begin{aligned}
&p(W, X, Y, Z, \mu, \Sigma, r^{in}, r^{out} | \pi, A, \Psi, \Phi, \theta, \Theta_0) \\
=&\prod_{s}^{S} p\left(Z_{s,1} | \pi\right) \cdot \prod_{s}^{S} \prod_{t=2}^{T} p\left(Z_{s,t} | Z_{s,t-1}, A\right) \\
&\cdot \prod_{s}^{S} \prod_{t=1}^{T} \prod_{n}^{N} p\left(W_{s,t,n}, Y_{s,t,n}, X_{s,t,n} | Z_{s,t}, \Psi, \Phi, \theta\right) \\
&\cdot \prod_{s}^{S} \prod_{t=1}^{T} p(c_{s,t}^{in} | b_{s,t}^{in} \cdot \lambda_l^{in}) \cdot p(c_{s,t}^{out} | b_{s,t}^{in} \cdot \lambda_l^{in}) \\
&\cdot p(\lambda_l^{in}, \lambda_l^{out} | \alpha^{in}, \beta^{in}, \alpha^{out}, \beta^{out})
\end{aligned}
\tag{3.11}
$$

Thus, searching for the best setting of the model parameters for STM-S is equivalent to the maximization of the logarithm of the joint distribution in Equation 3.11.

The time consumption consists of the above algorithm consists of two parts: 1) computation of the forward-backward algorithm; and 2) the computation of Equations A.1 $\sim$ A.19. The time complexity of the first part is $S \cdot T \cdot K$, where S is the number of sequences, T is the time length of a sequence, K is the number of latent states. The time complexity of the second part is $S \cdot T \cdot V \cdot K + S \cdot T \cdot V \cdot K \cdot J$, where N is the size of the vocabulary and J is the number of latent topics. Combine the two parts and multiply the number of EM iterations q, the comprehensive time complexity is $S \cdot T \cdot V \cdot K \cdot J \cdot q$.

Considering the large number of S, which is the number of all the historical sequences, the batch EM algorithm for estimating the model parameters is quite time-consuming. Moreover, as Twitter is streaming in real time, the batch-based updating of the model parameter given the newly-coming data requires the calculation of the whole historical training set, which is prohibitively expensive in practical usage. To solve this challenging, in this work, an online parameter optimization method is proposed, as introduced in the following section.

3.4.2 Online Parameter Optimization Algorithm

This section first proposes online parameter optimization algorithm for STM-I, and then presents that for STM-S.

Parameter optimization for STM-I

Unlike ordinary batch EM algorithm, being able to perform on-line estimation means that the data must be run through only once [24, 25]. The basic rationale of online EM algorithm is to replace the expectation step by a stochastic approximation step, while keeping the maximization step unchanged.

In this work, for STM-I, corresponding stochastic E-step is designed, including the computation of the conditional expectations. However, unlike batch algorithm where all the event-specific language models θ^R are optimized iteratively, θ^R for each newly-coming event in online algorithm cannot be known beforehand. Hence the likelihood in Equation 3.10 is unknown, which prevents the calculation of $\mathrm{E}[p(Z_{s_i,t} = k)]$.

To address this problem, this research work proposes to maximize the likelihood in Equation 3.10 with respect to θ^R, n^R, and n^B, which can be simplified into Equation 3.14.

After calculating θ^R, the conditional expectations of unknown parameters can be obtained by Stochastic E-step, which is elaborated in Appendix.

Utilizing the above stochastic E-step, the parameter of STM-I is trained on the fly of data stream, as summarized in Algorithm 1. Specifically, the current sequence of social media message s_i is crawled from data stream, which is utilized to calculate the conditional expectations for current data point by Steps 5, 7, 9, and 11. Then, the conditional expectations are used to update the sufficient statistics $\hat{\mu}_{l,k,i}$, $\hat{\Sigma}_{l,k,i}$, $g_{s,k,w,i}$, and $f_{k,j,w,i}$ in real time, as shown in Steps 6, 8, and 10. Finally, the maximum likelihoods of all the model parameters are calculated in Steps 13 $\sim$ 19. This EM iteration performs while the data is streaming until the end of the stream.

ALGORITHM 1: Online EM Algorithm for STM-I

Input: D, $\Theta_0 = \{\mu_0, \beta_0, \Lambda_0, \nu_0\}$
Output: A^*, π^*, Ψ^*, Φ^*, θ^*, μ^*, Σ^*.

Set the initial learning rate $\gamma_0 = 0.5$. Initialize θ^B, θ^R, Ψ, and Φ. Set $i = 0$;
repeat
  Get current sequence s_i from Twitter data stream;
  Obtain optimal θ^R by maximizing the likelihood in Equation 3.10;
  Calculate $E[p(Z_{s_i,t} = k)]$ using forward-backward algorithm;
  for $k \leftarrow 1$ **to** K **do**
    for $l \leftarrow 1$ **to** L **do**
      $\hat{N}_{l,k,i} \leftarrow (1-\gamma_i) \cdot \hat{N}_{l,k,i-1} + \gamma_i \cdot \sum_t \mathrm{E}[p(Z_{s_i,t} = k)]$;
      $\hat{\mu}_{l,k,i} \leftarrow (1-\gamma_i) \cdot \hat{\mu}_{l,k,i-1} + \gamma_i \frac{\sum_t \mathrm{E}[p(Z_{s_i,t}=k)](r^{in}_{s_i,t}, r^{out}_{s_i,t})}{\hat{N}_{l,k,i}}$;
      $E^{\hat{\Sigma}}_{l,k,i} \leftarrow \sum_t \mathrm{E}[p(Z_{s_i,t}=k)]\left(\hat{\mu}_{l,k} - (r^{in}_{s_i,t}, r^{out}_{s_i,t})\right)^2 / \hat{N}_{l,k,i}$;
      $\hat{\Sigma}_{l,k,i} \leftarrow (1-\gamma_i) \cdot \hat{\Sigma}_{l,k,i-1} + \gamma_i \cdot E^{\hat{\Sigma}}_i$;
      $\mu_{l,k,i} = (\beta_0 \mu_0 + \hat{N}_{l,k,i} \cdot \hat{\mu}_{l,k,i}) \Big/ (\beta_0 + \hat{N}_{l,k,i})$;
      $\Sigma_{l,k,i} = \frac{\Lambda_0 + \hat{\Sigma}_{l,k,i}}{\nu_0 + 3} + \frac{\beta_0 \hat{N}_{l,k,i} (\hat{\mu}_{l,k,i} - \mu_0)(\hat{\mu}_{l,k,i} - \mu_0)^T}{(\beta_0 + \hat{N}_{l,k,i})(\nu_0 + 3)}$;
    end
    for $j \leftarrow 1$ **to** J **do**
      for $w \leftarrow 1$ **to** V **do**
        $E^f_{k,j,w,i} \leftarrow \sum_t^T N_{s_i,t,w} \frac{\mathrm{E}[p(Z_{s_i,t}=k)] \cdot \Psi_{k,2} \cdot \Phi_{k,j} \theta^{B_j}_w}{\Psi_{k,1} \theta^R_{s_i,t,w} + \Psi_{k,2} \sum_j \Phi_{k,j} \theta^{B_j}_w}$;
        $f_{k,j,w,i} \leftarrow (1-\gamma_i) \cdot f_{k,j,w,i-1} + \gamma_i \cdot E^f_i$;
      end
      $\Phi_{k,j,i} = \sum_w f_{k,j,w,i} / \sum_j \sum_w f_{k,j,w,i}$;
    end
    for $w \leftarrow 1$ **to** V **do**
      $E^g_{s_i,k,w,i} \leftarrow \sum_t^T N_{s_i,t,w} \frac{\mathrm{E}[p(Z_{s_i,t}=k)] \cdot \Psi_{k,1} \cdot \theta^R_{s_i,t,w}}{\Psi_{k,1} \theta^R_{s_i,t,w} + \Psi_{k,2} \sum_j \Phi_{k,j} \theta^{B_j}_w}$;
      $g_{s_i,k,w,i} \leftarrow (1-\gamma_i) \cdot g_{s,k,w,i-1} + \gamma_i \cdot E^g_i$;
    end
    $\Psi_{k,1,i} = \sum_{s,w} g_{s,k,w,i} / (\sum_{s,w} g_{s,k,w,i} + \sum_{w,j} f_{k,j,w,i})$;
    $\Psi_{k,2,i} = \sum_w \sum_j f_{k,j,w,i} / (\sum_s \sum_w g_{s,k,w,i} + \sum_w \sum_j f_{k,j,w,i})$;
  end
  for $j \leftarrow 1$ **to** J **do**
    for $w \leftarrow 1$ **to** V **do**
      $\theta^{B_j}_{w,i} = \sum_k f_{k,j,w,i} / \sum_k \sum_w f_{k,j,w,i}$;
      $\theta^{B_j}_{w,i} = \sum_k f_{k,j,w,i} / \sum_k \sum_w f_{k,j,w,i}$;
    end
  end
  $i \leftarrow i + 1$;
until *the end of data stream*;

Parameter optimization for STM-S

Similar to STM-I, for STM-S, the calculations of the conditional expectations can be obtained through Stochastic E-step.

The model parameters of STM-S θ^B, θ^R, Ψ, and Φ need to be initialized. The initializations of them follows the same strategy for STM-I.

Utilizing the above-proposed stochastic E-step, the parameter of STM-S is trained on the fly of data stream, as summarized in Algorithm 2. Specifically, the current sequence of social media message s_i is crawled from data stream, which is utilized to calculate the conditional expectations for current data point by Steps 7, 9, 13, and 15. Then, the conditional expectations are used to update the sufficient statistics $\hat{\lambda}^{in}_{c,l,k,i}$, $\hat{\lambda}^{in}_{b,l,k,i}$, $\hat{\lambda}^{out}_{c,l,k,i}$, $\hat{\lambda}^{out}_{b,l,k,i}$, $g_{s,k,w,i}$, and $f_{k,j,w,i}$ in real time, as shown in Steps 5, 8, 10, 14, and 16. Finally, the maximum likelihoods of all the model parameters are calculated in Steps 11 and 17 $\sim$ 23. This EM iteration performs while the data is streaming until the end of the stream.

Time Complexity Analysis

As being deduced in Section 3.4.1, for the batch algorithm, the time complexity of is $S \cdot T \cdot V \cdot K \cdot J \cdot q$.

For the online algorithm, the time complexity of E-step is: $K \cdot T \cdot V \cdot J \cdot h$, the time complexity of M-step is $K \cdot T \cdot (L + J \cdot V + V) + J \cdot W$. Hence the total time complexity is: $(K \cdot T \cdot V \cdot J \cdot h + K \cdot T \cdot (L + J \cdot V)) \cdot q$.

Therefore, the time complexity of online algorithm is independent to S, the number of sequences in the training set, while is linear in h, the number of iterations to optimize θ^R, the language model for event-specific expressions.

3.5 Spatiotemporal Event Forecasting

In this section, the spatiotemporal event forecasting is formalized as a sequence classification problem based on the models proposed above, and an effective method for calculating the sequence likelihood is presented.

3.5.1 Sequence Classification.

Given a sequence of tweets, it is first necessary to identify whether the underlying development revealed by this sequence will lead to an event or not. These two possibilities each has a corresponding set of sequences and the two proposed models are trained based on

ALGORITHM 2: Online EM Algorithm for STM-S

Input: D, $\Theta_0 = \{\mu_0, \beta_0, \Lambda_0, \nu_0\}$
Output: A^*, π^*, Ψ^*, Φ^*, θ^*,λ^{in_*}, λ^{out_*}.

Set the initial learning rate $\gamma_0 = 0.5$. Initialize θ^B, θ^R, Ψ, and Φ. Set $i = 0$;
repeat
  Get current sequence s_i from Twitter data stream;
  Obtain optimal θ^R by maximizing the likelihood in Equation 3.11;
  Calculate $E[p(Z_{s_i,t} = k)]$ using forward-backward algorithm;
  for $k \leftarrow 1$ **to** K **do**
    for $l \leftarrow 1$ **to** L **do**
      $\hat{N}_{l,k,i} \leftarrow (1-\gamma_i) \cdot \hat{N}_{l,k,i-1} + \gamma_i \cdot \sum_t \mathrm{E}[p(Z_{s_i,t} = k)]$;
      for $m \in \{in, out\}$ **do**
        $E^{\hat{\lambda}^m_c}_{l,k,i} \leftarrow \sum_t c^m_{s_i,t} \cdot \mathrm{E}[p(Z_{s_i,t} = k)]/\hat{N}_{l,k,i}$;
        $\hat{\lambda}^m_{c,l,k,i} \leftarrow (1-\gamma_i) \cdot \hat{\lambda}^m_{c,l,k,i-1} + \gamma_i \cdot E^{\hat{\lambda}^m_c}_i$;
        $E^{\hat{\lambda}^m_b}_{l,k,i} \leftarrow \sum_t b^m_{s_i,t} \cdot \mathrm{E}[p(Z_{s_i,t} = k)]/\hat{N}_{l,k,i}$;
        $\hat{\lambda}^m_{b,l,k,i} \leftarrow (1-\gamma_i) \cdot \hat{\lambda}^m_{b,l,k,i-1} + \gamma_i \cdot E^{\hat{\lambda}^m_b}_i$;
        $\lambda^m_{k,l,i} = \frac{(\alpha^m - 1) + \hat{\lambda}^m_{c,k,l,i}}{\beta^m + \hat{\lambda}^m_{b,k,l,i}}$;
      end
    end
    for $j \leftarrow 1$ **to** J **do**
      for $w \leftarrow 1$ **to** V **do**
        $E^f_{k,j,w,i} \leftarrow \sum_t^T N_{s_i,t,w} \frac{\mathrm{E}[p(Z_{s_i,t}=k)] \cdot \Psi_{k,2} \cdot \Phi_{k,j} \theta^{B_j}_w}{\Psi_{k,1}\theta^R_{s_i,t,w} + \Psi_{k,2} \sum_j \Phi_{k,j} \theta^{B_j}_w}$;
        $f_{k,j,w,i} \leftarrow (1-\gamma_i) \cdot f_{k,j,w,i-1} + \gamma_i \cdot E^g_i$;
      end
      $\Phi_{k,j,i} = \sum_w f_{k,j,w,i} / \sum_j \sum_w f_{k,j,w,i}$;
    end
    for $w \leftarrow 1$ **to** V **do**
      $E^g_{s_i,k,w,i} \leftarrow \sum_t^T N_{s_i,t,w} \frac{\mathrm{E}[p(Z_{s_i,t}=k)] \cdot \Psi_{k,1} \cdot \theta^R_{s_i,t,w}}{\Psi_{k,1}\theta^R_{s_i,t,w} + \Psi_{k,2} \sum_j \Phi_{k,j} \theta^{B_j}_w}$;
      $g_{s_i,k,w,i} \leftarrow (1-\gamma_i) \cdot g_{s,k,w,i-1} + \gamma_i \cdot E^g_i$;
    end
    $\Psi_{k,1,i} = \sum_{s,w} g_{s,k,w,i} / (\sum_{s,w} g_{s,k,w,i} + \sum_{w,j} f_{k,j,w,i})$;
    $\Psi_{k,2,i} = \sum_w \sum_j f_{k,j,w,i} / (\sum_s \sum_w g_{s_i,k,w,i} + \sum_w \sum_j f_{k,j,w,i})$;
  end
  for $j \leftarrow 1$ **to** J **do**
    for $w \leftarrow 1$ **to** V **do**
      $\theta^{B_j}_{w,i} = \sum_k f_{k,j,w,i} / \sum_k \sum_w f_{k,j,w,i}$;
      $\theta^{B_j}_{w,i} = \sum_k f_{k,j,w,i} / \sum_k \sum_w f_{k,j,w,i}$;
    end
  end
  $i \leftarrow i + 1$;
until *the end of data stream*;

these sequences: one model characterizes the development process leading to an event, while the other one characterizes the process that does not lead to an event. For the prediction, an unknown sequence will be aligned with the model in each class. This sequence will be classified into the class corresponding to the higher alignment score.

Denote C_1 as the model trained for the class corresponding to the situation: "future event" while C_2 is the model corresponding to "no event". Denote e_1 as the cost of misclassifying the first class as the second class while e_2 is the cost of for misclassifying the second class as the first class. The spatiotemporal event forecasting problem can be formalized as follows: Given a newly-arriving sequence of tweets s in location l, if $p(C_1|s,l) > \varepsilon \cdot p(C_2|s,l)$, then a future event is deemed likely to happen; $p(C_1|s,l) \leq \varepsilon \cdot p(C_2|s,l)$, where $\varepsilon = e_1/e_2$ is the cost ratio.

According to the Bayesian rule, we have $p(C_i|s,l) = p(s|C_i) \cdot p(C_i|l)/p(s)$, $i = 1, 2$, where $p(C_1|l)$ denotes the prior probability that an event occurrs in location l; $p(C_2|l) = 1 - p(C_1|l)$ denotes the prior probability that no event occurs in location l; $p(s)$ is a constant and thus can be omitted. If the historical record for location l is not available, the above Bayesian decision rule is formalized as $p(C_i|s) = p(s|C_i) \cdot p(C_i)/p(s)$, $i = 1, 2$, where $p(C_1)$ is the overall prior probability of event occurrence in any location, while $p(C_2) = 1 - p(C_1)$ denotes the prior probability that no event occurs. Finally, the sequence likelihood $p(s|C_i)$ is calculated based on the method described in the next section.

3.5.2 Calculation of Sequence Likelihood.

In a standard HMM, dynamic programming methods such as the Viterbi algorithm [31] are typically utilized to calculate the likelihood of the a newly-arriving sequence by finding the most likely sequence of latent states. In the proposed model, however, the traditional Viterbi algorithm is not applicable because the proposed model needs to determine the optimal language models $\theta^R = \{\theta^R_{s,t}\}^{S,T}_{s,t}$ that represent the words exclusive to this newly-arriving sequence. The calculation of sequence likelihood based on the proposed new model involves identifying the most probable latent states and the parameter θ^R that maximize the probability $p(s|C_i)$:

$$p\left(s|C_i\right) = \max_{\{Z_t\}^T_t, \theta^R, n^R, n^B} \ln p(s, Z_1, \cdots, Z_T|C_i) \tag{3.12}$$

where $n^R = \{n^R_{s,t}\}^{S,T}_{s,t}$ is the number of words explained by the language model θ^R in sequence s at time step t. $n^B = \{n^{B_j}_{s,t}\}^{S,T,J}_{s,t,j}$ is the number of the words explained by different latent topics. By introducing the notation ω_t such that $\omega_t \equiv \ln p(s, Z_1, \cdots, Z_t|C_i)$, Equation 3.12 can be solved by recursively calculating the following equation:

$$\omega_t = \max_{\theta^R_{s,t}, n^R_{s,t}, n^B_{s,t}} \ln p(s_t|Z_t, C_i) + \max_{Z_{t-1}} \{\ln p\left(Z_t|Z_{t-1}\right) + \omega_{t-1}\} \tag{3.13}$$

with the initial iteration: $\omega_t = \max_{\theta^R_{s,1}, n^R_{s,1}, n^B_{s,1}} \ln p(s_1|Z_1, C_i) + \ln p(Z_1)$. The variables $\{Z_t\}_t^T$ can be solved via a standard max-sum algorithm.

Next we address the optimization problem: $\max_{\theta^R_{s,t}, n^R_{s,t}, n^B_{s,t}} \ln p(s_t|Z_t, C_i)$. By referring to Equation 3.10 and omitting the constant term, the problem can be formalized as the following maximization problem:

$$\begin{aligned} &\max_{\theta^R_{s,t}, n^R_{s,t}, n^B_{s,t}} \sum_i^V n^R_{s,t,i} \cdot \log \theta^R_{s,t,i} + \sum_i^V \sum_j^J n^{B_j}_{s,t,i} \cdot \log \theta_i^{B_j} \\ &s.t. \sum\nolimits_i^V \theta^R_{s,t,i} = 1, n^R_{s,t,w} + \sum\nolimits_j^J n^{B_j}_{s,t,w} = \xi_w, n^R_{s,t,w} \geq 0 \\ &n^{B_j}_{s,t,w} \geq 0, \sum\nolimits_i^V n^{B_j}_{s,t,i} = \xi \cdot \Psi_{k,2}\Phi_{k,j}, \sum\nolimits_i^V n^R_{s,t,i} = \xi \cdot \Psi^R_{k,1} \end{aligned} \tag{3.14}$$

where ξ denotes the number of words in sequence s at time step t, $k = Z_t$ is the current latent state in sequence s and V is the size of the vocabulary. The coupling between the variables $n^R_{s,t,i}$ and $\theta^R_{s,t,i}$ prevents a globally optimal solution to this problem, so Lagrangian multipliers are added to enforce the constraints. Setting the derivative w.r.t. $\theta^R_{s,t,i}$ to 0, we obtain:

$$\frac{n^R_{s,t,i}}{\theta^R_{s,t,i}} + \gamma = 0 \tag{3.15}$$

where γ is the Lagragian multiplier for the first equality constraint. By utilizing the first two equality constraints in Equation 3.14, we can derive:

$$\theta^R_{s,t,i} = \frac{n^R_{s,t,i}}{\xi \cdot \Psi^R_{k,1}} \tag{3.16}$$

Substituting Equation 3.16 into Equation 3.14, we get

$$\begin{aligned} &\max_{n^R_{s,t}, n^B_{s,t}} \sum_i^V n^R_{s,t,i} \cdot \log \frac{n^R_{s,t,i}}{\xi \cdot \Psi^R_{k,1}} + \sum_i^V \sum_j^J n^{B_j}_{s,t,i} \cdot \log \theta_i^{B_j} \\ &s.t. n^R_{s,t,w} + \sum\nolimits_j^J n^{B_j}_{s,t,w} = \xi_w, n^R_{s,t,w} \geq 0, n^{B_j}_{s,t,w} \geq 0, \\ &\sum\nolimits_i^V n^{B_j}_{s,t,i} = \xi \cdot \Psi_{k,2}\Phi_{k,j}, \sum\nolimits_i^V n^R_{s,t,i} = \xi \cdot \Psi^R_{k,1} \end{aligned} \tag{3.17}$$

Here, the objective function in Equation 3.17 is convex with respect to $n^R_{s,t}$ and $n^{B_j}_{s,t}$. Therefore, the global solution can be found by using a traditional numerical optimization method, such as the interior point method [72]. After $n^R_{s,t}$ and $n^{B_j}_{s,t}$ are optimized, $\theta^R_{s,t}$ can be calculated based on Equation 3.16. Finally, the maximization problem in Equation 3.12 is solved and thus the sequence likelihood can be calculated.

3.6 Experimental Evaluation

This section presents an experimental evaluation of the effectiveness and efficiency of the proposed approach based on comprehensive experiments on Twitter data from two different countries to forecast civil unrest events such as protests and strikes in Mexico, and flu outbreaks in the United States. All the experiments were conducted on a computer with a 2.6 GHz Intel i7 CPU and 16 GB RAM.

Table 3.2: Datasets and event labels

Dataset	Time Period	# Raw Tweets	# Processed Tweets	#Events
Civil unrest	2013-01-01 - 2013-06-01	32,459,668	57,856	726
Flu	2011-01-01 - 2013-12-31	8,627,664,399	2,252,436	102

3.6.1 Experiment Design.

This subsection presents the configuration of the datasets, the gold standard report for these event labels (as shown in Table 3.2), data processing, comparison methods, parameter settings, and performance metrics.

Datasets: For the analysis of civil unrest events forecasting, 10 percent of raw Twitter data in Mexico was collected through Datasift's Twitter collection engine from Jan 1, 2013 to Jun 1, 2013. The data from Jan 1, 2013 to Feb 28, 2013 was used as training, and the remaining was used for testing. For the analysis of flu forecasting, tweets containing at least one of 124 predefined flu-related keywords (e.g., "cold", "fever", and "cough") were collected during the period from Jan 1, 2011 to Dec 31, 2013 in the United States. The data from Jan 1, 2011 to Jan 1, 2013 was used for training, and the rest was used for testing.

Gold Standard Report of Event Labels: The civil unrest forecasting results were validated against a labeled set called Gold Standard Report (GSR) that was exclusively provided by MITRE (see [86] for more details). The GSR was organized by manually harvesting civil unrest events reports from the 10 most significant news outlets[1] in Mexico and the world, as ranked by International Media and Newspapers[2]. There were totally 726 events during Jan 1, 2013 to Jun 1, 2013. An example of a labeled GSR event is given by the tuple: (CITY = "Hermosillo", STATE = "Sonora", COUNTRY = "Mexico", DATE = "2013-01-20"). The forecasting results of flu outbreaks were validated against the flu statistics reported by the Centers for Disease Control and Prevention (CDC). CDC publishes weekly influenza-like illness (ILI) activity level within each state in the United States using the proportion of the

[1]They are La Jornada, Reforma, Milenio, the New York Times, the Guardian, the Wall Street Journal, the Washington Post, the International Herald Tribune, the Times of London, and Infolatam.

[2]International Media and Newspapers website. Available: http://www.4imn.com/. Accessed on Oct 1, 2014

outpatient visits to healthcare providers for ILI. There are 4 ILI activity levels: minimal, low, moderate, and high, where the level "high" corresponds to salient flu outbreak and is considered for forecasting. There were in total 102 events during Jan 1, 2011 to Dec 31, 2013. A example of CDC flu outbreak event is: (STATE = "Michigan", COUNTRY = "United States", WEEK = "01-06-2013 to 01-12-2013").

Gold Standard Report of Event Labels: The civil unrest forecasting results were validated against a labeled set known as the Gold Standard Report (GSR) that was exclusively provided by MITRE[3]. GSR was generated by manually collecting the civil unrest events reports from the 10 most significant news outlets[4] in Mexico and the world, as ranked by International Media and Newspapers[5]. A total of 726 events were identified from Jan 1, 2013 to Jun 1, 2013. An example of a labeled GSR event is: (CITY = "Hermosillo", STATE = "Sonora", COUNTRY = "Mexico", DATE = "2013-01-20"). The forecasting results for the flu outbreaks were validated against the influenza statistics reported by the Centers for Disease Control and Prevention (CDC). The CDC publishes weekly influenza-like illness (ILI) activity level within each state in the United States based on the proportion of outpatient visits to healthcare providers for ILI and classifies them according to 4 activity levels: minimal, low, moderate, and high, where the level "high" corresponds to a flu outbreak. There were in total 102 events during Jan 1, 2011 to Dec 31, 2013. A example of a CDC flu outbreak event is: (STATE = "Michigan", COUNTRY = "United States", WEEK = "01-06-2013 to 01-12-2013").

Data Preprocessing: For the first data set, three labelers collectively labeled 20,906 tweets in both English and Spanish during Jun, 2012 to Feb, 2013. After two had labeled all the tweets into positive (i.e., relevant to civil unrest) or negative, all the tweets where they disagreed were sent to the third labeler for final determination. Consequentially, the tweets were labeled into 6,793 positive and 14,113 negative, and the results used to train a linear SVM classifier. For the second data set, the labeled set in [62] was utilized, and used these to train a linear SVM to identify tweets relevant to the flu. Both SVMs were generated based on unigram features containing all the distinct words with frequencies greater than 20 in the individual datasets. The trained SVM classifiers extracted the tweets deemed relevant to civil unrest and flu from the respective datasets. The locations of the tweets were extracted from the geotags (coordinates and places). All the tweets without geotags were discarded.

Comparison Methods: There are 4 proposed approaches evaluated in this work. They are STM-I, STM-S, and online versions of them, which are STM-I (online) and STM-S (online). The proposed approaches were compared with four representative methods and one baseline method. The *Autoregressive exogenous model (ARX)* [2] assumes that for each separate location, the count of future events is dependent on both the count of historical event

[3]MITRE website. Available: http://www.mitre.org/. Accessed Oct 1, 2014.

[4]These outlets are La Jornada, Reforma, Milenio, the New York Times, the Guardian, the Wall Street Journal, the Washington Post, the International Herald Tribune, the Times of London, and Infolatam.

[5]International Media and Newspapers website. Available: http://www.4imn.com/. Accessed on Oct 1, 2014

and the tweet volume. When forecast, an output above "1" indicates that an event has occurred; otherwise no event is deemed to have occurred. *The linear regression (LinReg) model* [9, 21, 49, 78] assumes that for each separate location there is a linear relationship between tweet observations and event occurrences ("0" denotes nonoccurrence, "1" denotes occurrence). The input feature here is the volume of domain-related tweets. When forecasting, an output below "0.5" indicates no event; an output over "0.5" indicates that an event has occurred. *In the Logistic regression (LogReg) model* [99] event forecasting is treated as a classification problem. The input features are the proportions of latent topics extracted from the tweet texts coming from a specific location based on latent dirichlet allocation. The output is "0" if there is no event and "1", if there is one. The *Kernel density estimation-based logistic regression (KDE LogReg) model* [41] forecasts the event occurrence at a location by considering the historical event numbers and the tweet semantics. The set of input features is a combination of: 1) the historical event numbers spatially smoothed by KDE; and 2) the proportions of latent topics of tweet content. Finally, the *baseline* method considers the probability of historical event occurrence to be the probability of future event occurrence. Note that this baseline is also used as the prior in the proposed new approach.

Parameter Settings: Except for the baseline method, which does not require parameters, all the comparison methods were implemented based on the algorithms presented in the original papers. The strategies recommended by the authors were followed strictly to select features and estimated the model parameters via 10-fold cross-validation. The new method proposed here has several prior parameters and three tunable parameters. The four prior hyperparameters were set as follows: The historical prior ratio mean μ_0 was set as the mean of the domain-related tweet ratios in all the locations and in all the time steps; the prior scale matrix Λ_0 was set as an identity matrix; the number of prior measurements β_0 was set to be 1; and the degrees of freedom ν_0 to the dimension of the vector $\mu_{k,l}$. The three tunable parameters are the misclassification cost ratio ε, the number of latent topics J and the number of latent states K and these were set as 10, 5, and 4, respectively, based on 10-fold cross-validation.

Performance Metrics: Three main performance metrics are considered: precision, recall, and F1-score. The reported forecasting alerts are structured as tuples of (date, location), where "location" is defined at the city level for civil unrest events, and state level for flu outbreaks. A forecasting alert is matched to a true event if both the date and the location attributes are matched; otherwise, it is considered to be a false forecast. Note that because the time granularity of CDC flu outbreak labels is at week-level, it is considered as a match in time if the forecast date of an alert falls within the week of a true flu outbreak event.

3.6.2 Event Forecasting Results

Table 3.3 presents the comparison between the proposed 4 approaches and the 5 competing methods for the task of forecasting civil unrest and flu outbreak events.

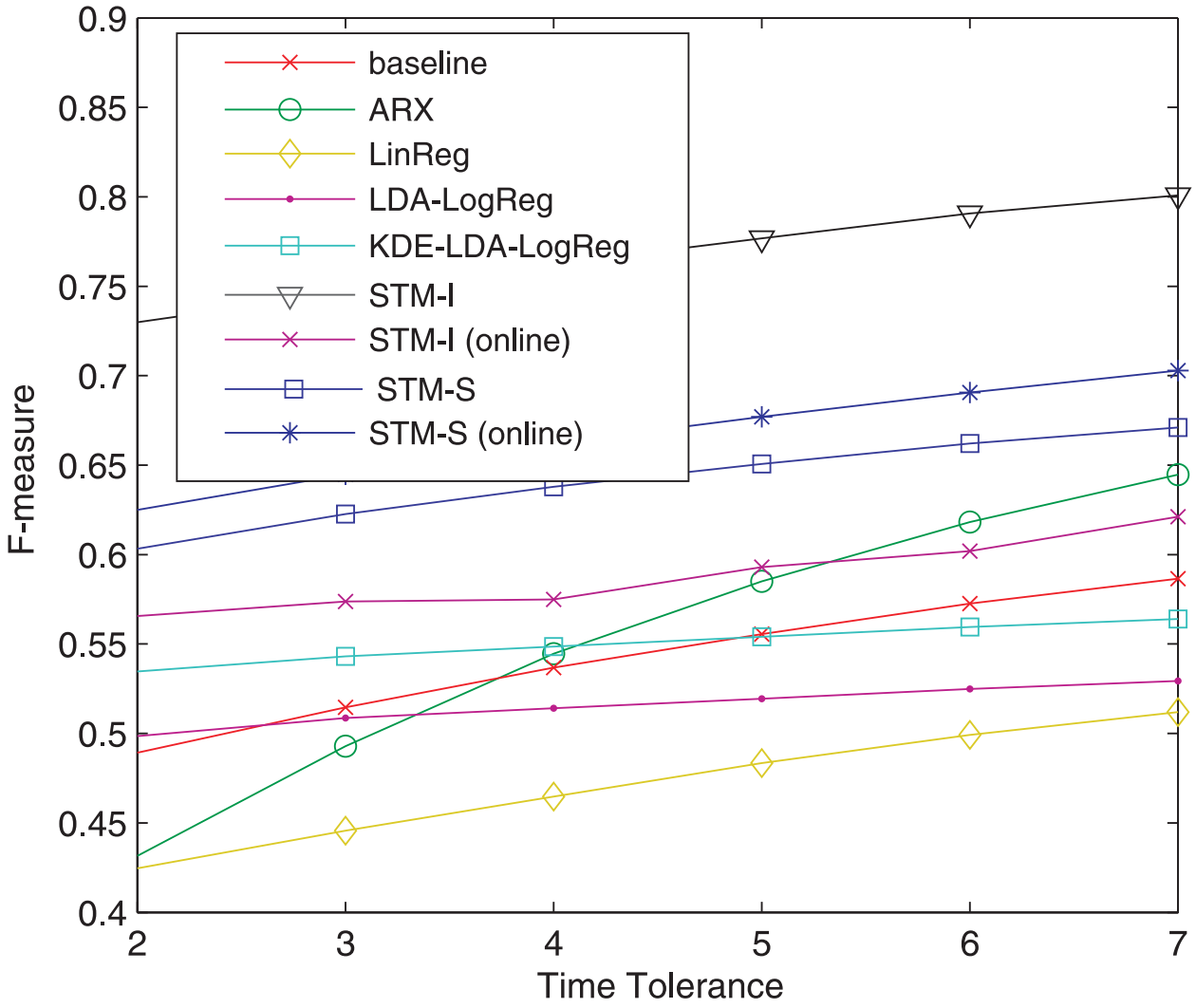


Figure 3.4: The prediction performance with respect to the tolerance of predicted time error on civil unrest dataset. The number of true positive is enlarged when the time tolerance increases.

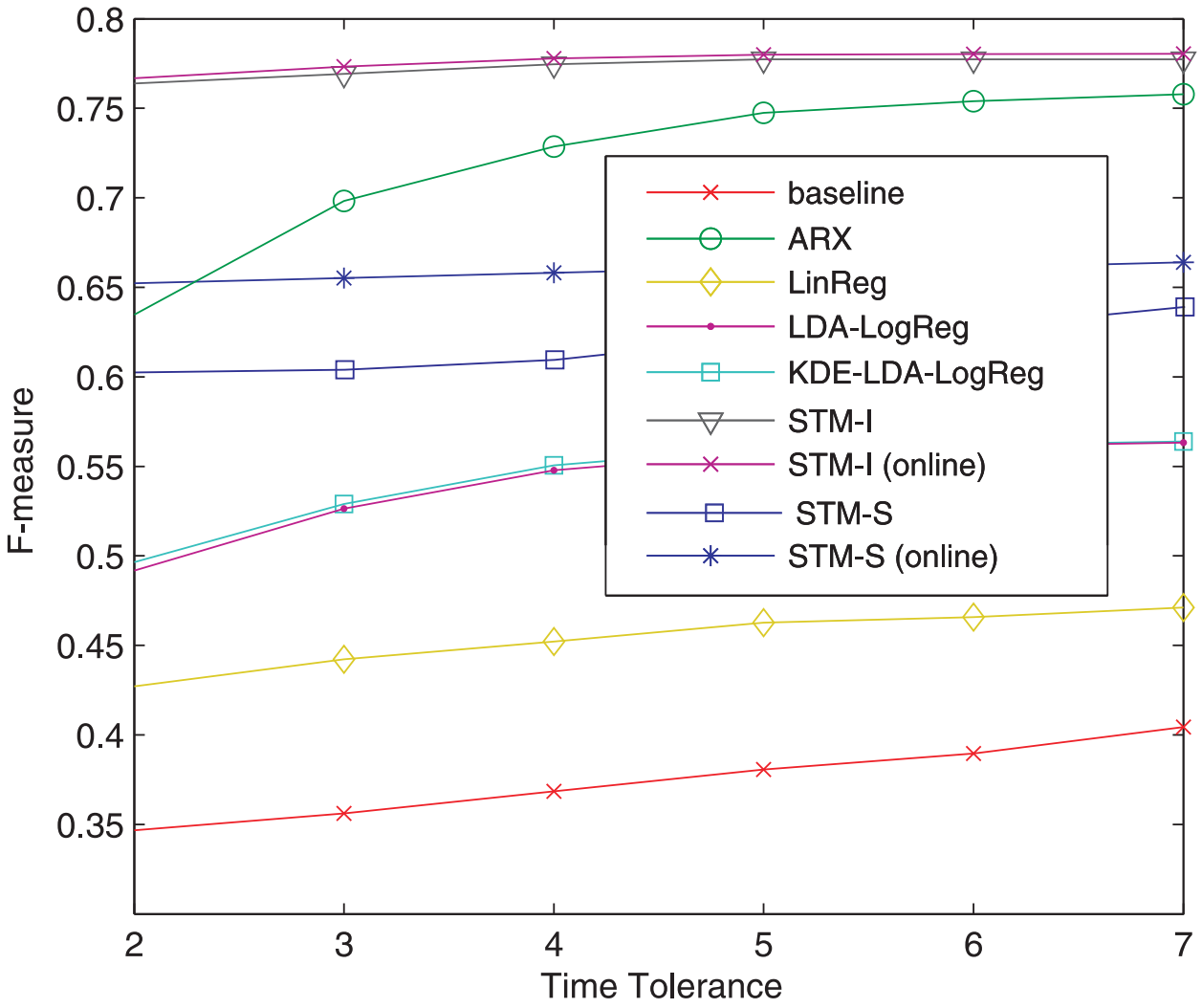


Figure 3.5: The prediction performance with respect to the tolerance of predicted time error on flus dataset. The number of true positive is enlarged when the time tolerance increases.

Table 3.3: Event forecasting results for the civil unrest

Metric	Precision	Recall	F-measure	Runtime
Baseline	0.44	0.59	0.50	0.001
ARX	0.26	0.43	0.32	0.001
LinReg	0.70	0.18	0.29	0.001
LDA-LogReg	0.31	0.70	0.43	0.005
KDE-LDA-LogReg	0.42	0.69	0.52	0.005
STM-I	0.75	0.70	0.72	0.32
STM-S	0.58	0.54	0.56	0.33
STM-I (online)	0.53	0.52	0.53	0.33
STM-S (online)	0.64	0.55	0.60	0.50

Table 3.4: Event forecasting results for the civil unrest and flu datasets

Metric	Precision	Recall	F-measure	Runtime
Baseline	0.28	0.39	0.33	0.001
ARX	0.14	0.66	0.23	0.001
LinReg	0.64	0.31	0.41	0.01
LDA-LogReg	0.27	0.55	0.36	0.02
KDE-LDA-LogReg	0.78	0.32	0.46	0.03
STM-I	0.83	0.69	0.75	2.1
STM-S	0.63	0.53	0.58	2.5
STM-I (online)	0.69	0.76	0.72	2.0
STM-S (online)	0.68	0.52	0.59	2.5

For the civil unrest dataset, the proposed new approaches achieved the best overall performance in precision, recall, and F1-score, outperforming the five comparison methods by up to 38% in F1-score and 7% in precision. This could be because the proposed approach considers the spatial burstiness as well as the tweet content, which is crucial for the forecasting of civil unrest events. Among the proposed new approaches, STM-I, which is Gaussian-distribution batch-based model, achieves the best performance. The batch-based approaches, STM-I and STM-S generally outperform the online-version of them in performance. But the newly proposed online version-based approaches still outperform competing methods by a substantial margin on both precision and recall. KDE Logistic Regression achieved a F1-score that was 21% higher than those of ARX, LinReg, and LogReg due to its consideration of spatial dependencies. The poor performances of ARX and LinReg indicate that focusing solely on tweet volume is insufficient for the task of civil unrest event forecasting. Thus, the tweet content as well as the spatial burstiness are important factors. The baseline method achieved a reasonably-well performance, indicating that it captured important historical event counts in different locations.

Table 3.4 demonstrates that the proposed approaches also consistently achieved the best performance in precision, recall, and F1-score, for the task of flu outbreak event forecasting. The F1-score of the proposed new approach was up to 63% higher than those of the five comparison methods. Among the proposed approaches, batch version of STM-I still achieves the best performance. Except the proposed approaches, KDE LogReg achieved the highest F1-score, suggesting the importance of considering spatial burstiness. The F1-score of the baseline was 34% lower than that in the civil unrest dataset, probably because the civil unrest events were clustered in several geographic regions, but the flu outbreak events were scattered across states. As a result, the use of prior information for event location distribution is effective in the civil unrest dataset, but noninformative in the flu data set. LinReg, on the other hand, achieved a 41% higher F1-score in the flu data set than in the civil unrest data set, which indicates that the tweet volume information plays an important role in this scenario. This could also explain why the comparison method LogReg, which only considers tweet semantics, achieved a poorer performance than in the civil unrest data set.

The proposed approaches and the five comparison methods all forecast next day events at the daily level. The running times of the proposed approach were on average 0.35 seconds per day on the civil unrest dataset, and 2.3 seconds per day on the flu dataset. These were markedly longer than the running time of the comparison methods for both datasets, primarily because the proposed approach considers the characterization of temporal correlations among tweet contents and the optimization of the language model for event-specific words. However, the running times achieved by the proposed approach were only a maximum of 3 seconds longer than those of the five comparison methods, and the resulting gain in forecasting accuracy of next day events makes this eminently practical for real-world applications.

In many application situations, it is not strictly required that the predicted time of events must be accurate within a timestep (e.g., date), but is acceptable to occur in next multiple time steps. For example, when doing civil unrest forecasting, people may be interested in predicting whether or not there will be event occurring in next multiple days. Instead of requiring very accurate predicted time, people sometimes may emphasize a sufficient lead time of forecasting. Similarly, when doing flu forecasting, people may be interested in forecasting whether or not the influenza activity will be high in next multiple weeks. To evaluate the performance of all the methods in this situation, the increase of correct predictions with respect to the increase of tolerance of predicted time error is validated in Figures 3.4 and 3.5.

In Figure 3.4, the F-measures of all the methods with respect to the tolerance of predicted time error are illustrated. It can be seen clearly that all the F-measures increase when the time tolerance increases. Among them, STM-I achieves the highest F-measure, about 0.80, when the time tolerance is 7 dates. ARX obtains a largest increasing rate, from 0.43 at 2 dates to 0.64 at 7 days, which indicates a robust prediction performance. STM-I and STM-I (online) also achieves competitive performance, around 0.70 when the tolerance time is almost 7 dates.

Figure 3.5 shows a similar pattern of Figure 3.4. First, the F-measures of all the methods increase when the tolerances of predicted time errors increase. Second, the methods STM-I, STM-I (online), STM-S, STM-S(online) and ARX are the best among all the methods. Third, the performance of ARX boosts up fastest when the time tolerance increases, finally achieves 0.76 F-measure at 7 dates.

3.6.3 Sensitivity Analysis

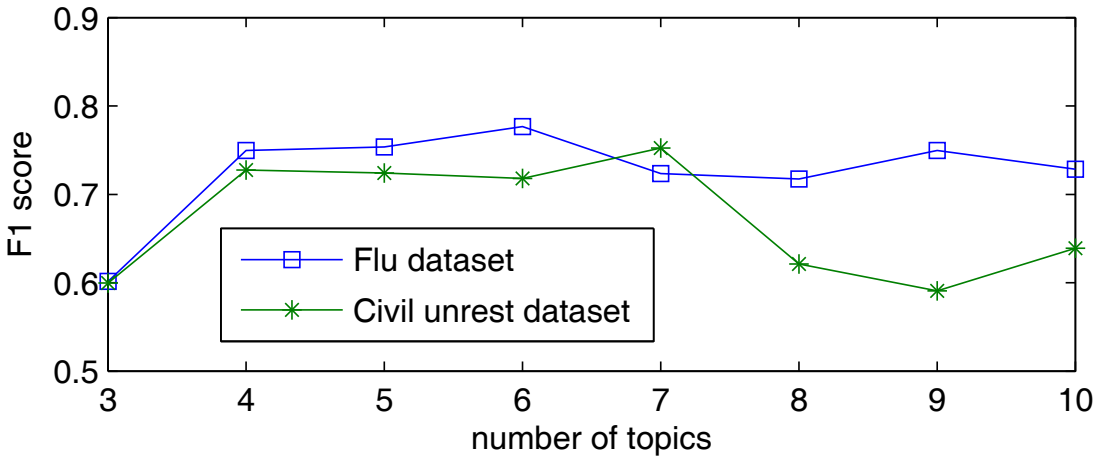


Figure 3.6: Sensitivity analysis on number of latent topics.

This section illustrates the sensitivity analysis. Here only the STM-I model is illustrated since the experiment results of the others follow the similar pattern.

Figures 3.6 and 3.7 illustrate the impact of the number of latent states and the number of latent topics on the event forecasting performance. By varying the number of latent topics from 3 to 10, the F1-score on the civil unrest and the flu data sets varies between 0.6 and 0.8. When the number of latent states was raised from 2 to 10, the perturbation in the F1-scores remained between 0.7 to 0.8 for both datasets. This indicates that the performance is less sensitive to the number of latent states than the latent topics in the given value interval of parameters. For both parameters, the performance for low values is relatively poor. For the number of latent topics, the range from 4 to 7 achieved the best performance, while for the number of latent states, the range from 4 to 9 corresponded to a good performance.

For both the civil unrest and flu datasets, the precision-recall curves of the new approach and the baseline method are shown in Figure 3.8(a) and Figure 3.8(b). To produce these curves, ε, the cost ratio of false positive to false negative was varied from 0.01 to 1 in increments of 0.01, and from 1 to 100 in increments of 1. For both civil unrest and flu forecasting, the performance of the proposed approach clearly outperformed the baseline.

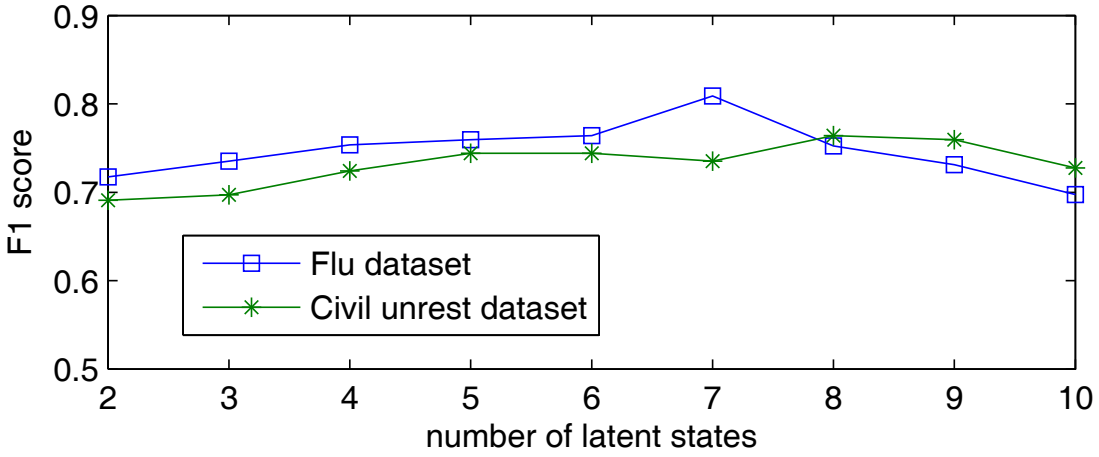


Figure 3.7: Sensitivity analysis on number of latent states.

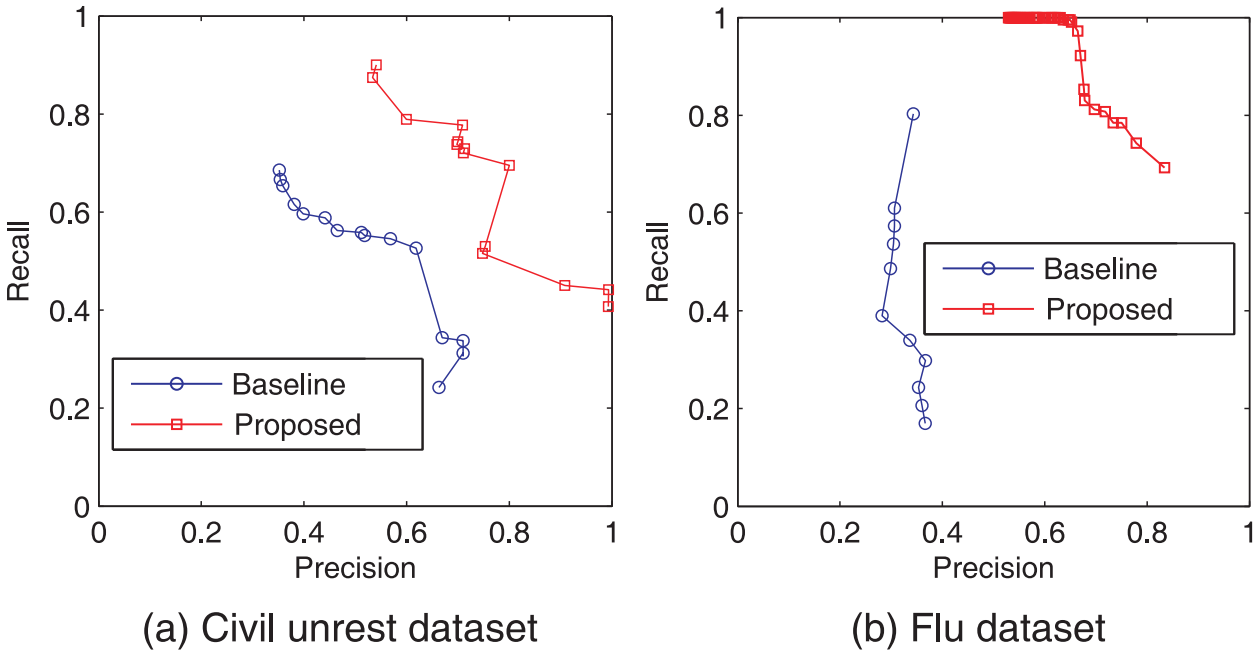


Figure 3.8: Precision-recall curves on civil unrest and flu data. The proposed model consistently outperforms the baseline when the cost ratio ε varies

3.6.4 Scalability

The training time of the batch-based models is typically sensitive to the size of training set. To validate this, Figure 3.9 and 3.10 illustrate the scalability on the number of training samples of the proposed 4 approaches on civil unrest dataset and flu dataset, respectively.

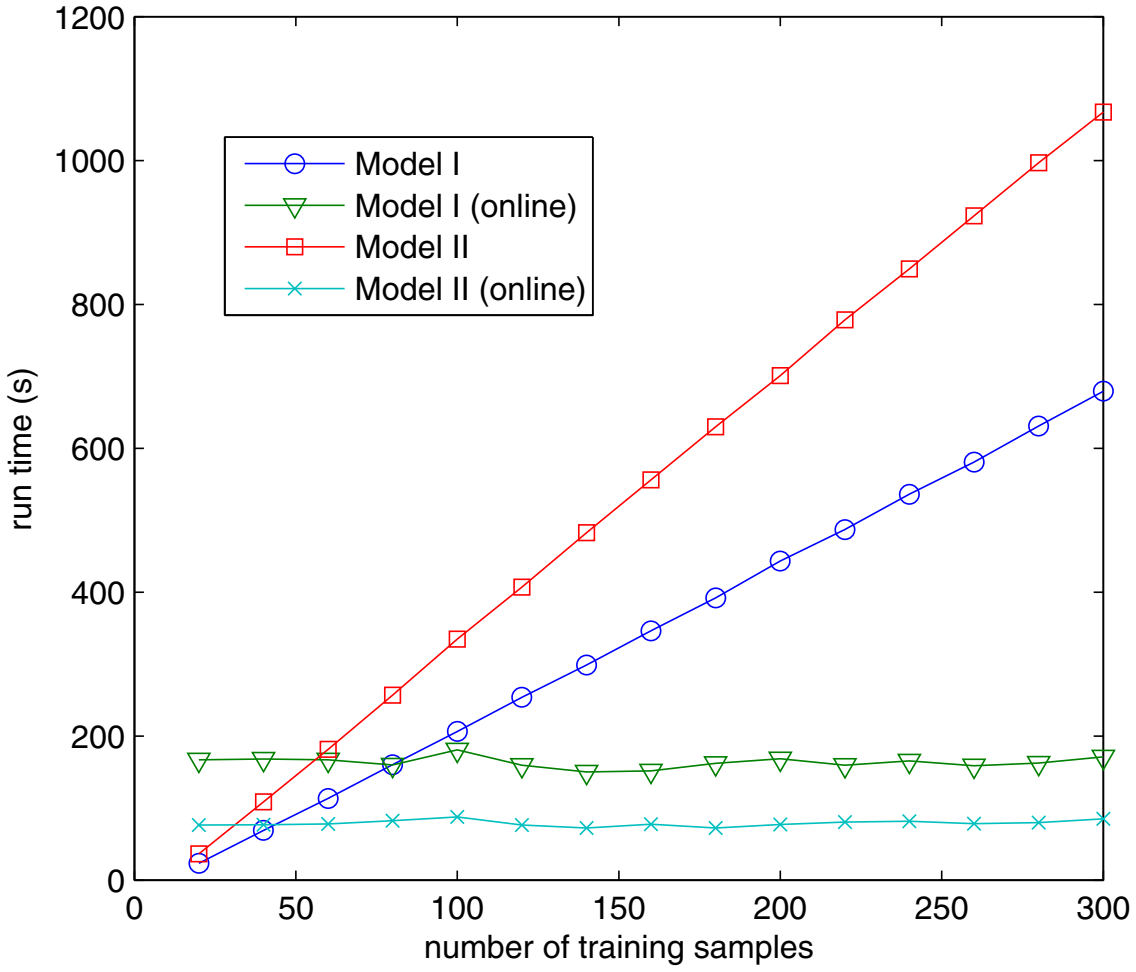


Figure 3.9: Scalability of the proposed models on civil unrest dataset. The runtime of batch-based models increase linearly with the size of training set while the runtime of online models is constant.

As shown in Figure 3.9, for civil unrest dataset, the run time for training STM-I and STM-S are linear in the number of training samples, starting from only 10 seconds with 20 samples until up to 1000 seconds with 300 samples. Different from these batch-based models, the training time of online versions: STM-I(online) and STM-S (online) are not sensitive to the number of training samples utilized, with a relatively constant run time around 150 seconds.

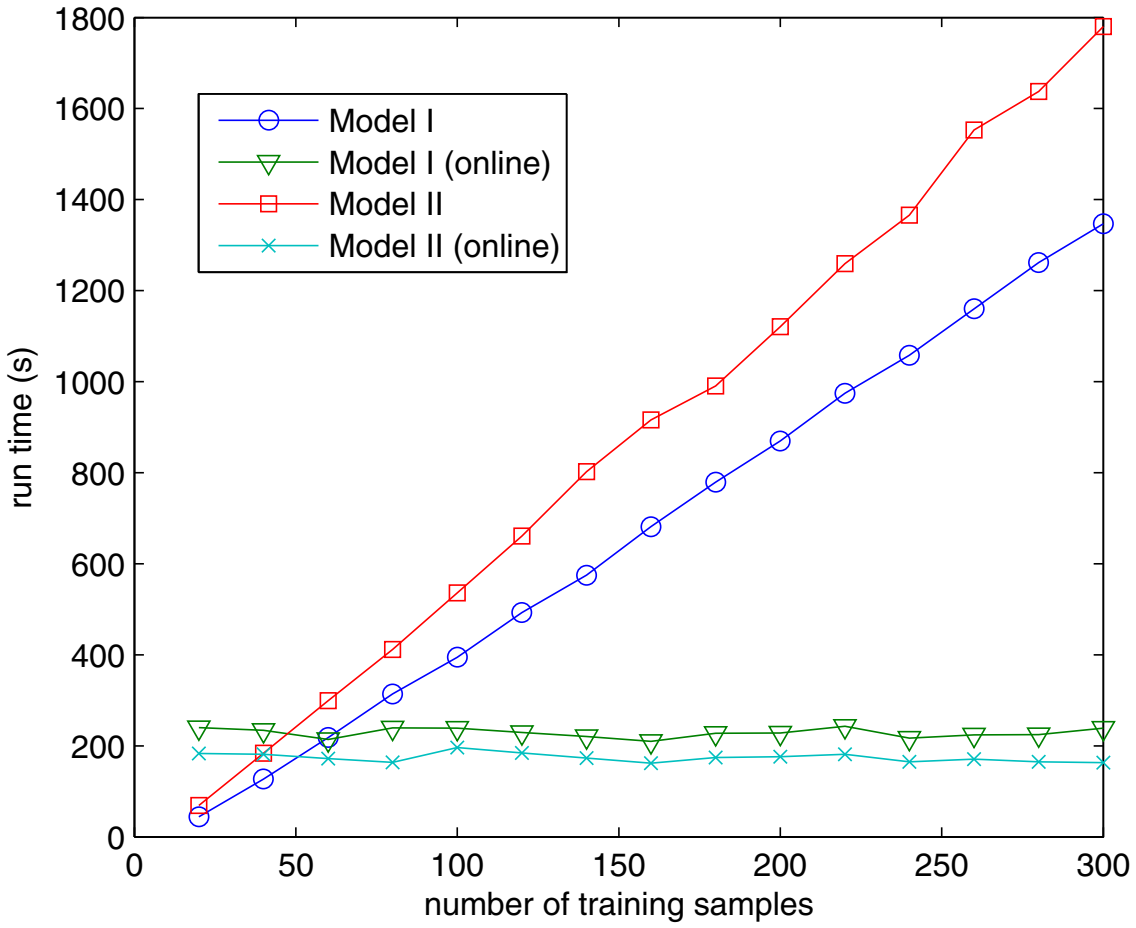
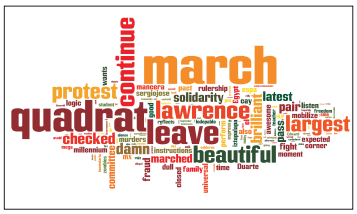


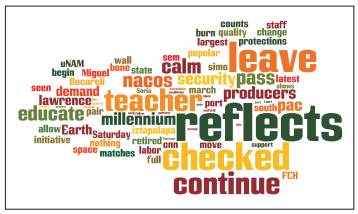
Figure 3.10: Scalability of the proposed models on flu dataset. The runtime of batch-based models increase linearly with the size of training set while the runtime of online models is constant.

On flu dataset, as shown in Figure 3.10, the run time for all the 4 approaches is larger than that on civil unrest dataset, due to the larger scale of the data. The run time for training STM-I and STM-S are, again, linear in the number of training samples, starting from only 10 seconds with 20 samples until up to 1600 seconds with 300 samples. The run time of the online versions of them is constantly around 200 seconds when the number of training samples varies from 20 to 300.

3.6.5 Case Study



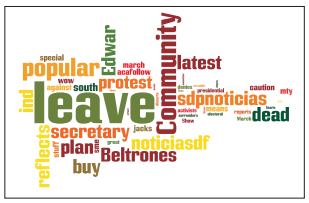
(a) topic 1 (describe protest)



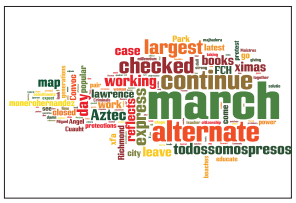
(b) topic 2 (background)

(c) topic 3 (describe protest)

(d) topic 4 (descriptive)



(e) topic 5 (background)



(f) topic 6 (descriptive)

(g) topic 7 (descriptive)

(h) topic 8 (call for protest)

(i) topic 9 (advertise protest)

(j) topic 10 (call for protest)

Figure 3.11: Illustration of all the 10 topics extracted (Translated in English). Topics 2 and 5 generally contain background words. Topics 1, 3, 4, 6, and 7 tends to include the descriptive words of the protests. Topics 8 and 10 focus on the words calling for protest. Topic 9 generally contains the words for disseminating planned protests.

Numerous interesting events predicted by the proposed approaches were observed. For instance, in this case study the forecasting of civil unrest event occurred on Mar 31th, 2013 using STM-I is shown. In the following, the topics, states, spatial burstiness, state transitions, event specific words are identified by the proposed STM-I, and is validated with real civil unrest events and flu outbreak events that are verified by authorized news outlets.

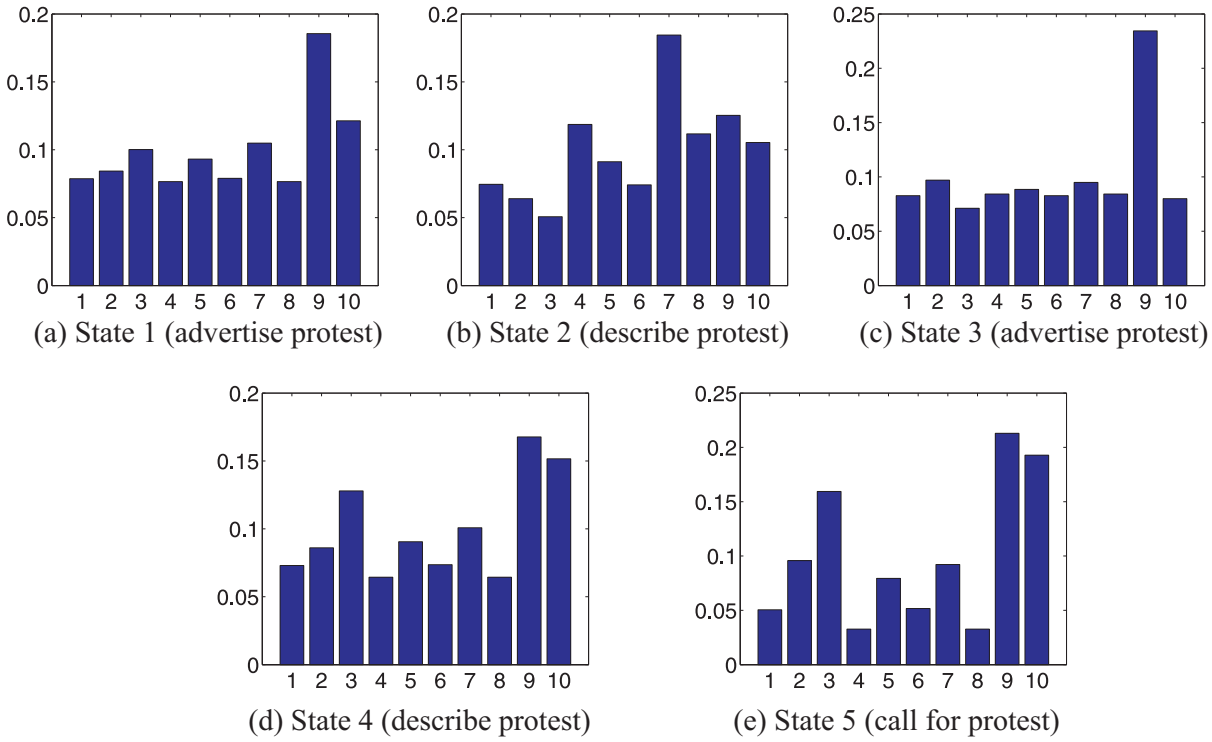


Figure 3.12: Contexts of the 5 latent states indicating the development stages of events. It can be seen that in different stages of the developing progression of protest, the distributions of topics are changing. States 2 and 4 could indicate the discussion about a protest among the public. States 1 and 3 could reveal the propaganda of the planed protests while State 5 might be related to the organization of the protest.

Case Study I: civil unrest event forecasting for Mexico on Mar 31th, 2013

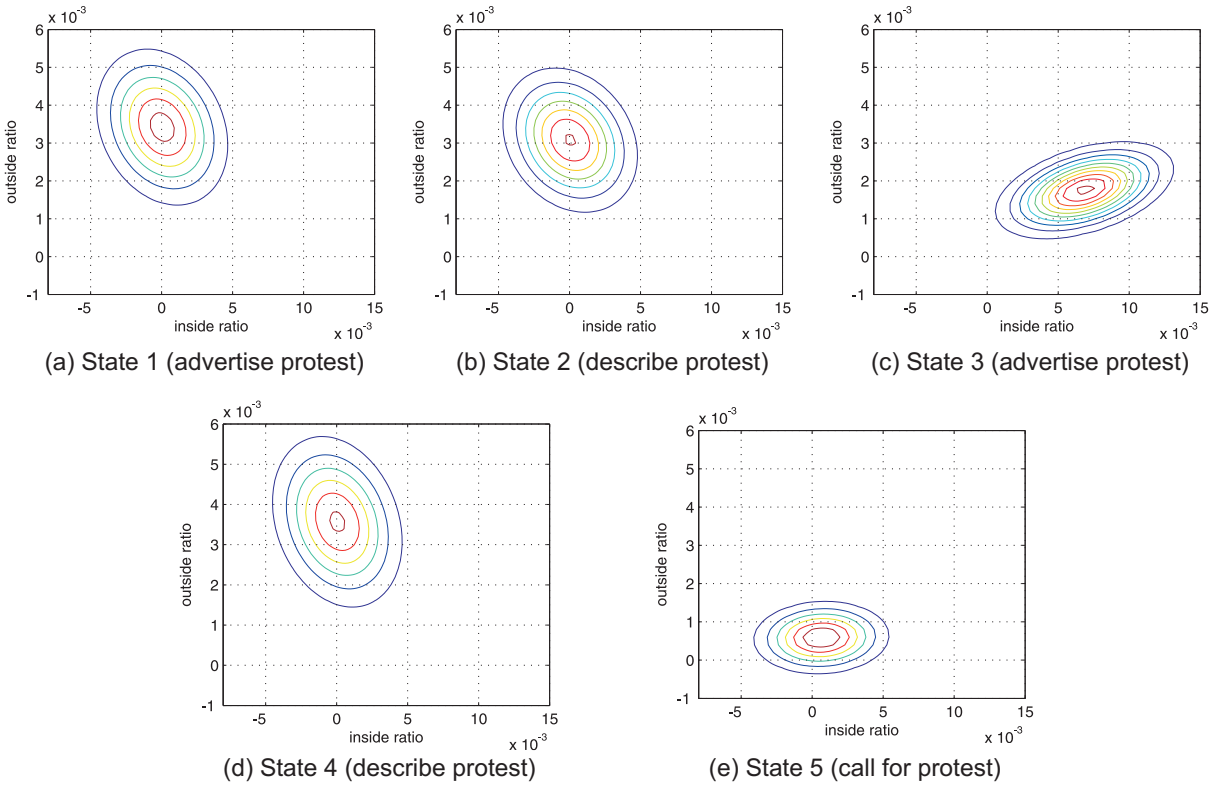


Figure 3.13: Spatial burstiness patterns of the 5 latent states indicating the development stages of events. States 1, 2, and 3 reveal the situation that the event-related tweets percentage inside the location is similar to that outside the location. States 3 and 5 shows that the event-related tweets percentage inside the location is much larger than that outside the location, which indicates a potential burstiness in the location.

Figure 3.11 illustrates the extracted topics for the civil unrest-related and common words of the proposed STM-I. It is ready to be seen that different topics explain the expressions about different stages of civil unrest. For example, Topic 1, 6, and 7, which highlights "march", "move", "plaza", and "takeover", generally talk about the description of a happened protest or planned protest. Topic 9 probably more concentrates on explaining the advertising of a planned protest, by the top keywords "largest", "problem", and "protest". And topic 8 and 10 explains the stage of "calling for protest", using the top keywords like "please", "call for", and "come on". Moreover, Topics 2 and 5 mainly absorbs some background common words such as "reflects", "continue", and "checked".

Figure 3.12 demonstrates the distribution of topics in each state of the proposed STM-I. For example, State 1 and State 3 tend to highlight the Topic 9 while weaken the influence from other topics. Therefore, State 1 is likely to indicate the dissemination of the planned protest. State 5 highlights the calling for protest because it leverages the topic 9 and 10. Differently,

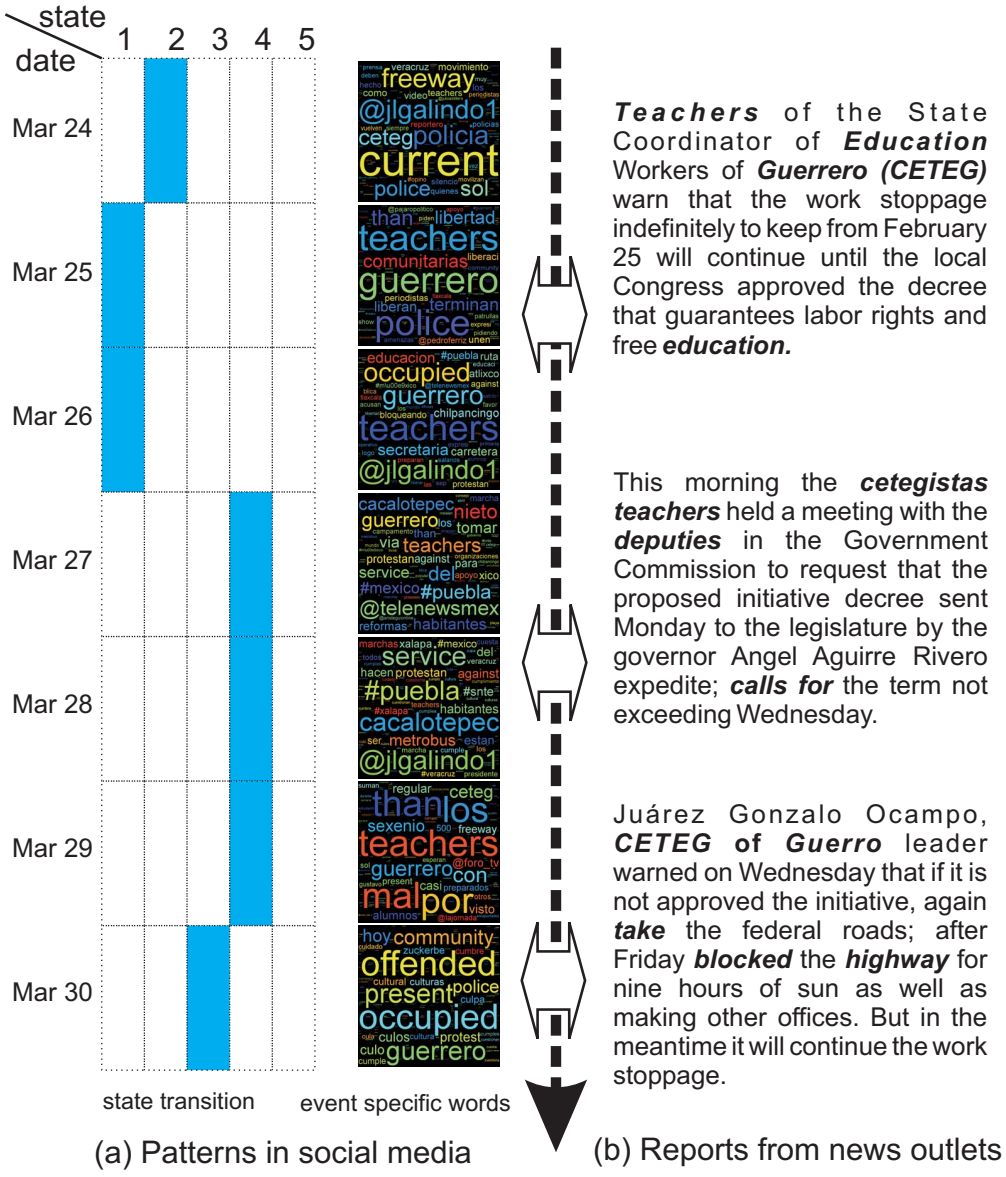


Figure 3.14: Event development progression discovered on microblogs are compared to the authorized reports by news outlets. The state transition on the left of (a) demonstrates the event stages conceptualized by the proposed model. On the right of (a), the word clouds shows that the keywords discovered in the microblogs match well with the bold keyword in the news reports in (b). The effective modeling of the development progression finally leads to accurate prediction of the occurrence of the events described in (c).

the most influential topic in State 2 is Topic 7, which indicates the emphasis of descriptions of the ongoing or past events.

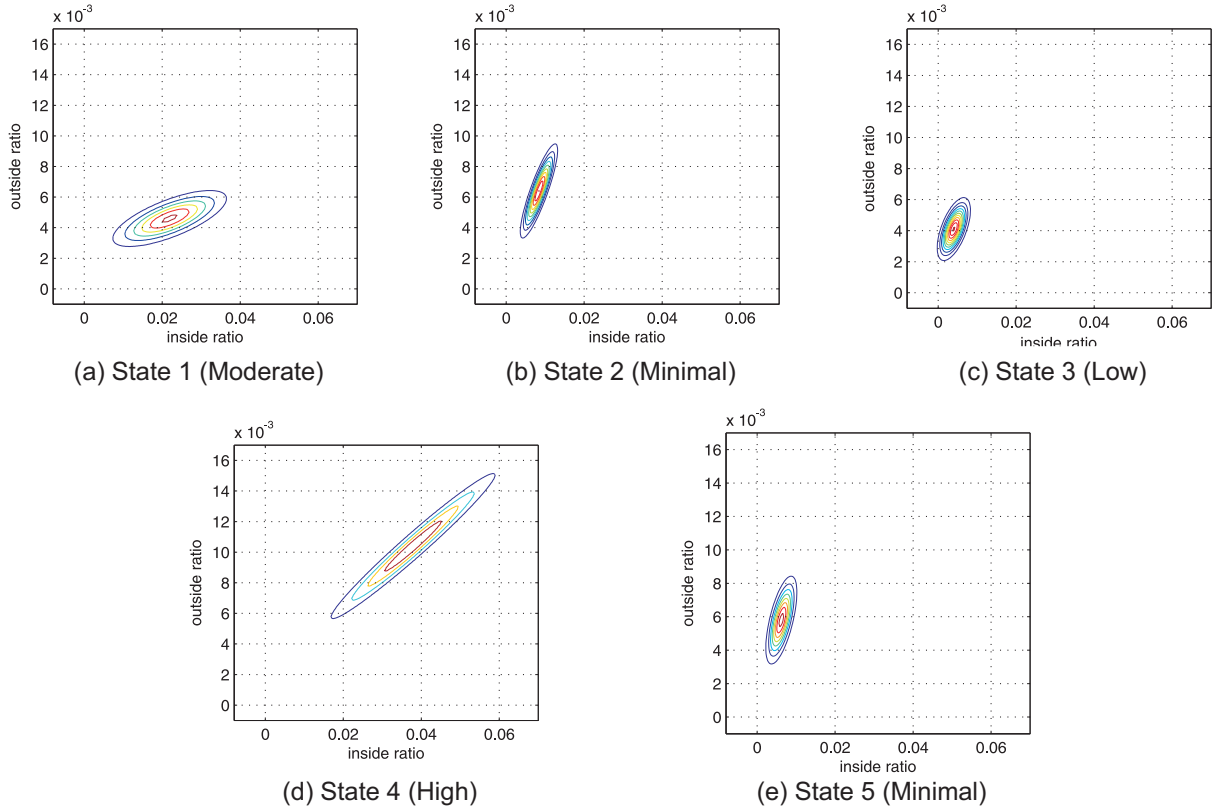


Figure 3.15: Spatial burstiness patterns of the 5 latent states indicating the development stages of flu events. States 2, 3, and 5 reveal the situation that the event-related tweets percentage inside the location is similar to that outside the location. States 1 and 4 shows that the event-related tweets percentage inside the location is much larger than that outside the location, which indicates a potential flu burstiness in the location.

Figure 3.15 shows the spatial burstiness in terms of inside and outside ratios for each state. To be specific, in each subplot, each state is illustrated as a bi-variate Gaussian whose means are the average inside and outside ratios of the location of current tweet sequence. And its variance reflects the degree how the ratios spread out and how the inside and outside ratios relate with each other. For example, State 1, 2, and 3 tend to be more similar with each other because their means of outside ratios are larger than those of their inside ratios. And the inside and outside ratios are likely to be negatively correlated as shown in Figure 3.15(a), 3.15(b), and 3.15(c). On the other hand, the States 3 and 5 are more likely to have larger inside ratios than outside ratios. And the inside and outside ratios are basically positively correlated. This generally indicates that there is burstiness occurred inside the location.

The development progress of an event (as described in Figure 3.14(c)) is reflected as the transitions among hidden states identified by the proposed STM-I as shown in Figure 3.14(a).

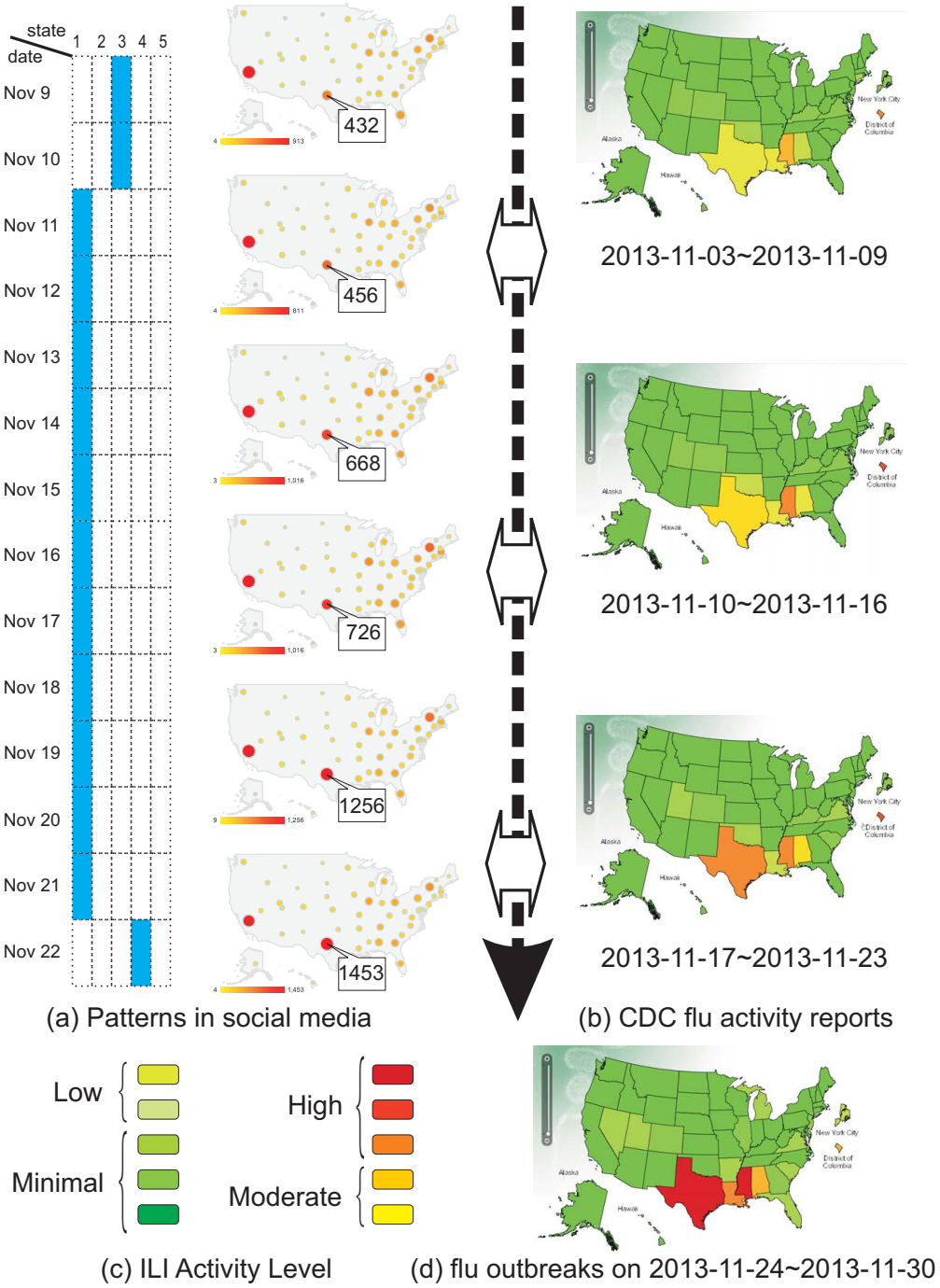


Figure 3.16: Flu event development progression discovered on microblogs are compared to the authorized reports by CDC. The state transition on the left of (a) demonstrates the event stages conceptualized by the proposed model. On the right of (a), the map shows that the increase of flu-related tweets in Texas match well with fast upgrading of reported flu activity in Texas as shown in (b). The effective modeling of the development progression finally leads to accurate prediction of the occurrence of the flu outbreaks illustrated in (c).

This progress is validated by the ground-truth descriptions from news reports, as shown in Figure 3.14(b).

As shown in Figure 3.14(a), the state transition is: State 2 → State 1 → State 1 → State 4 → State 4 → State 4 → State 3. By referring to Figures 3.11, 3.12, and 3.15, this state transition indicates that there were a potential development progress of "planning → advertising → calling". Figure 3.14(a) also illustrates the identified event-specific words for each date. Figure 3.14(b) demonstrates the identified event-specific words match the ground-truth from the news reports, especially for the keywords like "Guerrero" (protest location and protest target), "teacher" (protest initiator), and "occupy" (protest action). Therefore, the case study demonstrates that the topics, states, spatial-burstiness, and the state transitions identified by the proposed approach are effective, accurate and with practical meanings, which match the ground truth from the authorized news outlets.

Case Study II: flu outbreak event forecasting for Texas, USA on Nov 24th - Nov 30th, 2013

Figure 3.15 shows the spatial burstiness in terms of the inside and outside ratios for each latent state. To be specific, similar to Figure 3.15, in each subplot, a state is illustrated as a bi-variate Gaussian whose means are the average inside and outside ratios of the location of current tweet sequence. And its variance reflects the degree how the ratios spread out and how the inside and outside ratios relate with each other. For example, State 2, 3, and 5 tend to be more similar with each other because their means of outside ratios are larger than those of their inside ratios. Low inside and outside ratios indicates low activity of influenza. The inside and outside ratios are likely to be positively correlated as shown in Figure 3.15(b), 3.15(c), and 3.15(e). On the other hand, the States 1 and 4 are more likely to have larger inside ratios than outside ratios. The much larger inside ratio indicates a strong flu-related signal in social media from the current location, which generally reveals that there is or will be burstiness occurring inside the location.

The left part of Figure 3.16(a) shows the latent state transition while 3.16(b) shows the flu-related tweets distribution across the whole country. The initial latent state is State 3 on Nov 9 and Nov 10, which then transfers to State 1 on Nov 11-Nov 21 and finally goes to State 4 on Nov 22. By referring to Figure 3.15, we know that and State 3 indicates a moderate inside ratio and low outside ratio while State 1 indicates a high inside ratio and low outside ratio. By modeling this state transition and the flu-related tweets spatial distribution, the proposed model forecasts that there were a potential development progress of "flu outbreaks" in Texas in the following week. Figure 3.14(b) shows the flu activity level identified by the authorities, namely the CDC flu reports. It clearly demonstrates the upgrading of the flu activity in Texas from 2013-11-03 to 2013-11-23, and achieved a flu outbreak on the following week, which is consistent with the pattern that is identified and modeled through social media.

3.7 Conclusion

This work presents a novel model for spatiotemporal event forecasting in Twitter. The new generative approach uncovers the underlying development of events by jointly considering the structural semantics and the spatiotemporal burstiness of Twitter streams. Both batch and online-based inference algorithms are developed to optimize the model parameters. Utilizing the trained model, the alignment likelihood of tweet sequences is calculated by dynamic programming. Extensive empirical testing demonstrated the effectiveness of the new approach by comparing it with five representative methods. For future work, this approach is expected to be extended to other applications, such as forecasting other disease outbreaks and local events such as road congestion.

Chapter 4

Multi-Task Learning for Spatiotemporal Event Forecasting

This chapter proposes a novel multi-task learning framework which aims to concurrently address all the challenges. Section 4.1 introduces the research background. Section 4.2 reviews background and related work, and Section 4.3 introduces the problem setup. Section 4.4 presents new multi-task feature learning models, and Section 4.5 presents efficient algorithms based on IGHT. Experiments on real Twitter datasets are presented in Section 4.6, and the paper concludes with a summary of the research in Section 4.7.

4.1 Introduction

Microblogs such as Twitter and Weibo are experiencing an explosive level of growth. Millions of worldwide microblog users broadcast their daily observations on an enormous variety of topics, e.g., crime, sports, and politics.

This work focuses on the problem of spatial event forecasting from microblogs, for events such as civil unrest, disease outbreaks, and crime hotspots. The basic idea is to search for subtle patterns in specific cities as indicators of ongoing or future events, where each pattern is defined as a burst of context features (keywords) relevant to a specific event. For instance, the expression of discontent about gas price increases could be a potential precursor to a protest about government policies.

There are three technical challenges in addressing this problem: 1) **Dynamic features**. The language used in microblogs is highly informal, ungrammatical, and dynamic. Most existing methods treat fixed keywords as features [90, 92]. However, the expression in tweets may dynamically evolve, which makes the use of fixed features and historical training data insufficient. For example, the most significant Twitter keyword for the Mexican protests in

Aug 2012 was "#YoSoy132" (i.e., the hashtag of an organization protesting against electoral fraud), alluding to the protests against the Mexican presidential election, but "#CNTE" (i.e., a hashtag denoting the national teacher's association of Mexico) has become the most popular term by the beginning of 2013 due to the movements against the Mexican education reform. Ideally an event forecasting system must combine judicious use of static (fixed) features but must be cognizant to subtle changes involving dynamic features. 2) **Geographic heterogeneity.** Different cities have different characteristics, such as population, weather (e.g., humidity, temperature), and administrative structures (e.g., capital cities versus non-capital cities). As a result, it is difficult to impute basal levels of occurrence uniformly. Considering civil unrest as an example, finding 1000 tweets mentioning the keyword "protest" is likely not a strong indicator of an upcoming civil unrest event if the city houses a population of a few million users but could be a strong signal for a city with a population of 10,000. At the same time, it is difficult to dynamically adjust such thresholds precisely due to the data sparsity problems in the latter case. 3) **Scalability.** The massive scale of microblogging data necessitates development of new, scalable forecasting methods.

In order to concurrently address all these technical challenges, this work presents a novel computational approach in the framework of multi-task learning (MTL) that combines the strengths of methods that use static features (e.g., LASSO regression [86]) and those that use dynamic features (e.g., dynamic query expansion (DQE) [109]). these methods have been utilized, individually, for event forecasting and this work tackles challenges involved in unifying these contrasting approaches in a single framework. Learning multiple related tasks simultaneously effectively increases the sample size for each city, which can potentially improve the forecasting performance, especially when the sample size for each task (city) is small. One critical issue in multi-task learning is how to define and exploit the commonality among different tasks. Intuitively, events that occur around the same time may involve similar topics, and therefore tweets from different cities may share many common keywords that are related to the event(s). This chapter presents three multi-task feature learning (MTFL) formulations for event forecasting that differ in the specifics of how common features are extracted.

The main contributions of this study are summarized as follows:

1. **Formulation of a multi-task learning framework for event forecasting.** This chapter formulates event forecasting for multiple cities in the same country as a multi-task learning problem. In the proposed model, event forecasting models are built for different cities simultaneously by restricting all cities to select a common set of features. Both penalized and constrained MTL formulations are explored, which use different strategies to control the common set of features selected.

2. **Concurrent modeling of static and dynamic terms.** The existing models (LASSO and DQE) use different but complementary information; LASSO uses static terms, while DQE identifies dynamic terms. The newly proposed MTL formulations make use of both types of information by integrating the strengths of LASSO (a supervised

approach) and DQE (an unsupervised approach). To the best of our knowledge, there is not much prior work that combines supervised and unsupervised approaches for event forecasting.

3. **Development of efficient algorithms.** This chapter then explores both convex and non-convex optimization formulations. For convex problems, proximal methods, e.g., FISTA [13] are leveraged, which have been shown to be efficient for solving sparse and multi-task learning problems. For non-convex problems, the iterative Group Hard Thresholding (IGHT) [20] framework are employed, which is guaranteed to converge to a local solution.

4. **Comprehensive experiments to validate the effectiveness and efficiency of the proposed techniques.** The proposed methods are evaluated using Twitter data collected from July 2012 to May 2013 in 4 countries in Latin America: Mexico, Brazil, Paraguay, and Venezuela. For comparison a broad range of other algorithms are implemented. Results showed that the proposed methods consistently outperformed competing methods, including LASSO, DQE, traditional multitask learning models, and their variants. sensitivity analysis are also evaluated to reveal the impact of the parameters on the performance of the proposed methods.

4.2 Related Work

Compared to traditional media, Twitter has the following significant characteristics: 1) Timeliness of messages: Unlike traditional media that take hours or days to publish, tweets can be posted instantly utilizing portable mobile devices; 2) Ubiquity of social sensors: Tweets reflect the public's mood and trends, which could be the determinants of future social events; and 3) Availability of geo-information: Twitter users provide rich location information in profiles, texts, and geotags. As a social "sensor" which can identify emerging patterns in sentiments and opinions, the use of microblogs holds great promise for detection and forecasting of significant societal events.

The typical dichotomy to event detection or forecasting research is to classify them into whether they are supervised or unsupervised. The former consider a set of stationary terms whose distribution can be learned from historical data. Particularly, LASSO regression methods estimate a sparse predictive model based on a predefined set of keyword terms (vocabulary) for each city that predicts the probability of an ongoing event in this city in each predefined time interval (e.g., hourly or daily) [86]. Burst detection methods search for geographic regions (cities) where the aggregated counts of some predefined terms are abnormally high compared with the counts outside the cities. For example, Sakaki et al. consider spatiotemporal Kalman filtering, which is similar to space-time burst detection, to track the geographical trajectory of hot spots of tweets related to earthquakes [90]. Unsupervised methods, as the name indicates, consider a set of dynamic terms that could be different in

different time intervals, and apply unsupervised learning techniques for event detection. Particularly, the dynamic query expansion method (DQE) iteratively expand a predefined set of seed terms (e.g., protest, strike, march) using the current tweets to identify and rank new terms that are relevant to ongoing events, then retain the top terms and tweets containing these terms for further modeling [109]. Clustering-based methods search for novel spatial clusters of documents or terms using predefined similarity metrics, such as cosine similarity and social similarity for documents [4], or auto-correlations [2] and co-occurrences [100] for terms.

Event detection: A large body of work focuses on the identification of ongoing events, including earthquakes [90], disease outbreaks [92], and other types of events [4, 54, 63, 100]. In general, they either use classification or clustering to extract tweets of interest and examine the spatial [90], temporal [91, 100], or spatiotemporal burstiness [63] of the extracted tweets. However, instead of forecasting events in the future, these approaches typically can only uncover them after their occurrence.

Event forecasting: Most research in this area focuses on temporal events and ignores the underlying geographical information, such as the forecasting of elections [78,96], stock market movements [21], disease outbreaks [2,89], box office ticket sales [9,110], and crimes [99]. These works can be grouped into three categorizes: 1) Linear regression models: Simple features, such as tweet volumes, are utilized to predict the occurrence time of future events [9,21,49,78]; 2) Nonlinear models: More sophisticated features such as topic-related keywords are used as the input to build forecasting models using existing methods such as support vector machines or LASSO [89, 99]; 3) Time series-based methods: Methods like autoregressive models are used to model the temporal evolution of event-related indicators (e.g., tweet volume) [2]. However, there are few existing approaches that can provide true spatiotemporal resolution to predicted events. In [41], Gerber utilized a logistic regression model for spatiotemporal events forecasting using topic-related tweet volumes as features. Wang et al. [98] developed a spatiotemporal generalized additive model to characterize and predict spatio-temporal criminal incidents, but their model requires the demographic data. Ramakrishnan et al. [86] built separate LASSO models for different locations to predict the occurrence of civil unrest events. Zhao et al. [55, 86, 109] designed a new query expansion method to expand both keywords and key tweets by considering both semantic and social network relationships, and used the burstiness of key tweets to predict civil unrest events. Zhao et al. [110] designed a new predictive model based on topic model that jointly characterizes the temporal evolution in both semantics and geographical burstiness of social media content.

Multi-task learning: Multi-task learning (MTL) learns multiple related tasks simultaneously to improve generalization performance [27, 93]. Many MTL approaches have been proposed in the past [113]. In [39], Evgeniou et al. proposed the regularized MTL which constrained the models of all tasks to be close to each other. The task relatedness can also

be modeled by constraining multiple tasks to share a common underlying structure, e.g., a common set of features [8], or a common subspace [7]. MTL approaches have been applied in many domains, including computer vision and biomedical informatics. To our best knowledge, ours is the first work that applies MTL for civil unrest forecasting.

4.3 Problem Setup

Suppose there are m locations (e.g., cities, states) in a country of interest, and each location l has $n_{l,t} \in \mathbb{Z}$ tweets in each time interval t (e.g., hour, day). Define a matrix $C_{l,t} \in \mathbb{Z}^{p \times n_{l,t}}$, whose (i,j)-th entry, denoted as $C_{l,t,i,j}$, refers to the frequency of the i-th term in the j-th tweet. Here p refers to the size of the vocabulary V. We are also given a binary variable $Y_{t,l} \in \{0,1\}$ for each location l at time t, which indicates the occurrence ('yes' or 'no') of a future event. The goal is to predict the occurrence of a future event for a specific location l at a specific time interval t based on the tweets data collected.

This work is built upon two of our previous predictive models, including LASSO [86] and dynamic query expansion (DQE) [109]. Suppose we have a predefined subset of keywords of size d in V that are relevant to the event of interest for forecasting, and denote A as the corresponding incidence matrix, $A \in [0,1]^{d \times p}$. Define a matrix $K_{l,t}$ as follows: $K_{l,t} = A \cdot C_{l,t} \cdot \mathbf{1}$, where $\mathbf{1}$ refers to a vector of all ones. It is clear that $K_{l,t} \in \mathbb{Z}^{d \times 1}$ is the vector of keywords frequencies in location l at time t. The LASSO model learns a separate sparse linear regression model for each location l:

$$\arg\min_{w_l} \left\| {w_l}^T K_{t,l} - Y_{t,l} \right\|_2^2 + \rho_1 \|w_l\|_1,$$

where the regularization parameter ρ_1 controls the sparsity, and $w_l \in \mathbb{R}^{d \times 1}$ is the vector of regression coefficients that need to be estimated. We need to estimate $m \cdot d$ parameters in total for the m separate LASSO regression models.

DQE is a Twitter-oriented query expansion method to get dynamic keywords, which are then utilized for event detection or forecasting. Denote $I(\cdot)$ as the indicator function. For each location l and time t, define the number of tweets containing any of the k dynamic keywords $S_t^{(k)}$ as $D_{l,t,k}$. Then, the DQE-based event forecasting can be formulated as a function $Y_{l,t} = I(D_{l,t} > \gamma)$, that is, $Y_{l,t} = 1$ if $D_{l,t}$ is larger than the threshold γ; $Y_{l,t} = 0$, otherwise. The dynamic keywords are expanded and ranked from the seed query based on the tweets data C_t, where the seed query S_0 is an initial set of few semantically coherent keywords that characterize the concept of the targeted domain. Specifically, the keyword expansion process is formulated as follows:

$$P_t = F_t(B_t^T \cdot B_t + B_t^T R_t B_t) \cdot P_0$$

where $P_0 \in \mathbb{R}^{|V| \times 1}$ is the initial weight vector of all the words in V, $[P_0]_{i,1} = I(V_i \in S_0)$, and V_i is the ith word. B_t is the adjacency matrix between tweets and words. $R \in \mathbb{R}^{|C_t| \times |C_t|}$ is

the tweet-replying matrix, i.e., $[R_t]_{ij} = 1$ means there is replying relationship between tweet i and tweet j; $[R_t]_{ij} = 0$, otherwise. $F \in \mathbb{R}^{|V|\times|V|}$ is the inverse document frequency (IDF) matrix of F, which is a diagonal matrix such that $[F]_{ii}$ refers to the IDF of the word V_i. $P_t \in \mathbb{R}^{|V|\times 1}$ is the updated weight vector. Finally, the dynamic keyword set $S_t^{(k)}$ is defined as the top k words with the largest weights according to P_t.

There are three main challenges for using either of LASSO and DQE individually: (1) The LASSO model only uses a set of predefined fixed keywords, called "static features," which may not capture the fast-evolving expressions in Twitter, thus it may be difficult to predict future events that are related to a small set of new keywords not included in the fixed keywords set. (2) The LASSO model trains an individual model for each location, but many small cities may have insufficient amount of information in the training set to build an accurate forecasting model. (3) DQE requires two types of thresholds, which are 1) k, the number of dynamic keywords expanded from a seed query, and 2) γ, the least number of tweets, each of which contains any of dynamic keywords, to indicate the event occurrence. However, it is difficult to set these two thresholds based on domain experience. Next section presents a novel computational approach based on multi-task learning to address all these three challenges.

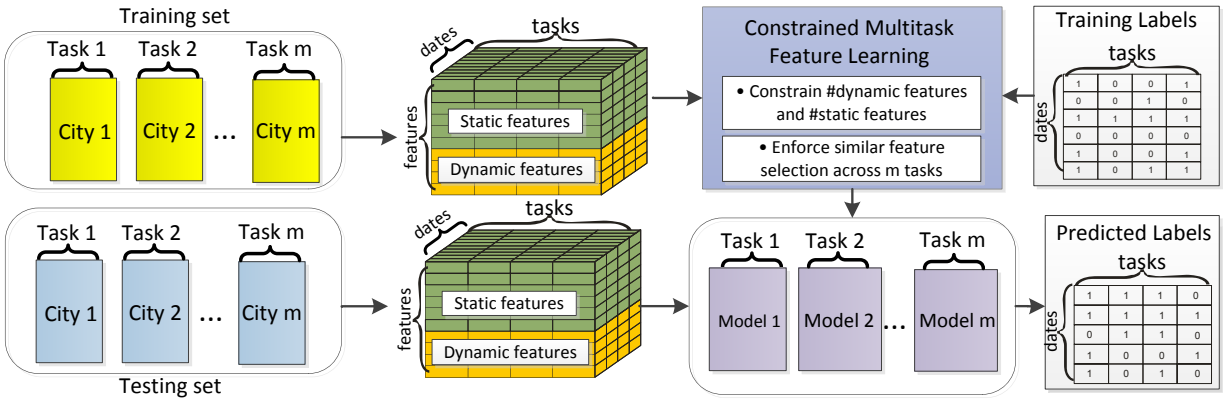


Figure 4.1: The flowchart of the proposed multi-task learning model

4.4 Models

As defined above, LASSO uses the "static feature" set $K_{l,t}$, which is the count of predefined keywords in location l at time t. DQE uses the "dynamic feature" set $D_{l,t,k}$, which is the number of tweets containing top k dynamic keywords at location l at time t. Because it is difficult to predefine an optimal k, this work proposes to make use of multiple k values in the range of $[1, s]$ (here s is user-specified parameter; the experiments show that using a set of $s = 20$ values is sufficient), and then learn the optimal k automatically in the proposed

multi-task learning framework. This results in $D_{l,t} = \{D_{l,t,k}\}_{k=1}^{s}, D_{l,t} \in \mathbb{R}^{K\times 1}$, called the "dynamic feature" set for location l and time t. The information used in LASSO and DQE are combined by forming a new data matrix $X_{l,t} = [K_{l,t}; D_{l,t}] \in \mathbb{R}^{d+s\times n_{l,t}}$. For notational simplicity, subscript t will be removed throughout the rest of this work.

This work aims to build m models $\{w_i|i = 1, \ldots, m\}$ to predict the occurrence of events for the m locations. A simple approach is to learn these m models (tasks) independently, ignoring the task relatedness. However, such approach does not consider the intrinsic relationships among cities, and the resulting models may not be accurate as some cities may not have sufficient information in the training set. To address this issue, this work aims to build the forecasting models for all m cities simultaneously by extracting and utilizing appropriate shared information across tasks [113]. Figure 4.1 illustrates the proposed multi-task learning framework. Learning multiple related tasks simultaneously effectively increases the sample size for each city, since when we learn a model for a specific city, we use information from all other cities.

Intuitively, the events that occur at different cities around the same time may involve similar topics, thus the tweets from different cities may share many common keywords that are related to the events. This motivates us to explore multi-task feature learning (MTFL) models which constrain multiple related models to select a common set of features. Specifically,three multi-task feature learning models are explored:

- Regularized multi-task feature learning model,
- Constrained multi-task feature learning model I,
- Constrained multi-task feature learning model II.

Each of the three models formulates the multi-task learning problem by following a general paradigm, i.e., to minimize a penalized empirical loss:

$$\min_W f(W) + \lambda g(W) \tag{4.1}$$

or a constrained version:

$$\min_W f(W) \text{ s.t.} \quad g(W) \leq l. \tag{4.2}$$

where $f(W)$ is the empirical loss on the training set; a smooth and convex loss function is used, e.g., the least squares and logistics loss. $g(W)$ is the regularization term that encodes task relatedness, which is typically non-smooth or even non-convex. λ (or l) is a tuning parameter to balance the tradeoff between the loss and penalty.

Different regularization/constraint terms capture different types of task relatedness [1,32,39, 53]. In this work, the least square loss is adopted, and the model relatedness are characterized by restricting all models to select a common set of features.The three models are detailed below.

4.4.1 Regularized MTFL Model

The j-th element in model w_i indicates the importance of j-th feature for i-th task. In MTFL, all tasks are restricted to share a common set of top features, that is, the forecasting models for all cities are based on the same subset of features. This can be achieved by grouping the j-th elements of all tasks together and selecting the top groups. Specifically, we consider the m entries of the j-th row of the matrix W as a group and use the $l_{2,1}$-norm regularization to identify the top groups [8]. Thus, the j-th feature which corresponds to the j-th element in models are likely to be selected or not by all models simultaneously, achieving our desired goal. Mathematically, the following multi-task feature learning model is employed:

$$\min_{W} \sum_{i=1}^{m} \|w_i^T X_i - Y_i\|_F^2 + \rho_1 \|W\|_{2,1} + \rho_{L2} \|W\|_F^2, \tag{4.3}$$

where the first term is the data fitting term for all tasks, $\|W\|_{2,1}$ denotes the $l_{2,1}$ norm of matrix W which encourages all tasks to select a common set of features, and it can be computed as the summation of l_2-norm of each row in W. The regularization parameter ρ_1 controls the sparsity. a small multiple of the Frobenius-norm regularization is included, i.e., $\|W\|_F^2$, to enhance the robustness of the model. Problem (4.3) is a convex problem and can be solved by the FISTA algorithm [13].

4.4.2 Constrained MTFL Model I

In the regularized MTFL model above, the model sparsity is controlled by the parameter ρ_1, which is less interpretable than the number of features selected. It is thus desired to develop a model which directly controls the number of features to be selected. To this end, this chapter introduces a constraint in the model which ensures that a specific number of rows of W will be non-zero, i.e., the number of features included in the model is controled. In particular, the following constrained multi-task feature learning model is considered:

$$\begin{aligned} &\min_{W} \sum_{i=1}^{m} \|w_i^T X_i - Y_i\|_F^2 + \rho_1 \|W\|_F^2, \\ &\text{s.t. } \sum_{j} I(\|w^j\| > 0) \leq r. \end{aligned} \tag{4.4}$$

Here w^j is the j-th row of W and $I(\cdot)$ is the indicator function. The constraint in (4.4) ensures that the number of nonzero rows of W is no larger than r, ensuring no more than r features will be selected. Note that the convexity property does not hold any more for Model (4.4). The iterative Group Hard Thresholding framework to solve (4.4) is utilized. More details are provided in the next section.

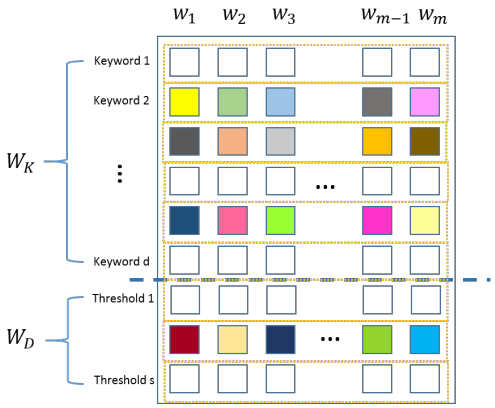


Figure 4.2: Illustration of constraint MTFL model II. Each column represents the model for a specific city. The i-th row in W_K indicates the feature values for the i-th static feature (i.e., keyword), and the j-th row in W_D corresponds to the j-th dynamic feature (i.e., threshold value). Colored entries represent non-zero values in the model matrix, while white entries represent zeros.

4.4.3 Constrained MTFL Model II

The constrained model above does not distinguish the static and dynamic features. Recall that the first d features correspond to the d static features, while the last s features correspond to the use of s dynamic features. The feature values thus have very different meanings. In general, d is much larger than s. In the experiments, d is around 2000, while s is around 10 to 20. Thus, it is desired to restrict the number of features selected from these two groups separately. In addition, in the current DQE model, only one dynamic feature is used and a common threshold value is applied for all cities in the same country. It is thus natural to restrict the number of dynamic features selected (out of the total s candidates) to be one. To achieve these goals, the following model is proposed, which selects u features from the d static features, and v features from the s dynamic features are selected:

$$\begin{aligned} \min_{W} & \sum_{i=1}^{m} \|w_i^T X_i - Y_i\|_F^2 + \rho_1 \|W\|_F^2, \\ \text{s.t.} & \sum_{j} I(\|w_K^j\| > 0) \leq u, \\ & \sum_{j} I(\|w_D^j\| > 0) \leq v, \end{aligned} \tag{4.5}$$

where W_K is the model matrix corresponding to the set of static features, and W_D is the model matrix corresponding to the set of dynamic features. The structure of the model is illustrated in Figure 4.2. Similar to Problem (4.4), u and v are user-specified parameters that control the number of features selected for the two sets of features, i.e., static feature

set and dynamic feature set, respectively. We set $v = 1$ in the experiments, however, the proposed model is more general in that the user can select an arbitrary number of dynamic features.

Problem (4.5) is non-convex due to the use of nonconvex constraints. Similar to Problem (4.4), the Iterative Group Hard Thresholding algorithm is applied to solve Problem (4.5). The details of the proposed algorithm for Problem (4.5) are shown in the next section.

4.5 Algorithm

The FISTA algorithm performs well for convex problems [13, 32, 113]. However, both Problem (4.4) and Problem (4.5) are non-convex. Even worse, they both involve discrete constraints, which make the problems challenging to solve. Motivated by the success of the iterative hard thresholding algorithm for solving l_0-regularized problems [20] and the recent advances on nonconvex iterative shrinkage algorithm [43,104], we propose to employ the Iterative Group Hard Thresholding framework to solve both problems. Note that Problem (4.4) is a special case of Problem (4.5) with $v = 0$. This chapter thus focuses on Problem (4.5) only in the following discussion. The details are summarized in Algorithm 1.

Algorithm 1 The Proposed Algorithm

Require: $\boldsymbol{X}$, $\boldsymbol{Y}$, ρ, $\eta > 1$
Ensure: solution $\boldsymbol{W}$

```
Initialize W^0, α^0 ← 1.
for i ← 1, 2, ... do do
    Initialize L
    repeat
        S^i ← W^i − (1/L)∇f(W^i)
        W^i ← proj(S^i) (defined in Lemma 1)
        L ← ηL
    until line search criterion is satisfied
    if the objective stop criterion satisfied then
        return W^i
    end if
end for
```

Recall Problem (4.4), and denote $f(W) = \sum_{i=1}^{m} \|w_i^T X_i - Y_i\|_F^2 + \rho_1 \|W\|_F^2$. The key idea of IGHT is to first use the gradient information at the current iterate to provide the first-order approximation of the objective function, then apply the projection operators to ensure the next iterate satisfies the given constraints. Specifically, we use the combination of the linear approximation of the function $f(W)$ at a given point W^0 and a quadratic penalty term, and

solve the following problem:

$$\begin{aligned} &\min_{W} f(W^0) + \langle \nabla f(W^0), W - W^0 \rangle + \frac{\rho}{2}\|W - W^0\|_F^2, \\ &\text{s.t. } \sum_j I(\|w_K^j\| > 0) \leq u, \\ &\quad \sum_j I(\|w_D^j\| > 0) \leq v, \end{aligned} \tag{4.6}$$

where ρ is a positive constant that can be estimated by a line search scheme. By ignoring the constants and re-arranging the terms in Problem (4.6), the following sub-problem is obtained:

$$\begin{aligned} &\min_{W} \frac{1}{2}\|W - S\|_2^2 \\ &\text{s.t.} \sum_j I(\|w_K^j\| > 0) \leq u \\ &\quad \sum_j I(\|w_D^j\| > 0) \leq v. \end{aligned} \tag{4.7}$$

where $S = W^0 - \frac{1}{c}\nabla f(W^0)$. Problem (4.7) aims to find the optimal point satisfying the constraint set that is closet to a fixed point S. It is called an Euclidean projection problem, denoted as proj$(\cdot)$, even the constraint set is not convex. The key of the IGHT framework is to solve the projection problem in (4.7). It is not hard to show that Problem (4.7) admits a closed-form solution as it can be decomposed into two independent problems, one for each block of features, as summarized in the following lemma.

Lemma 1. *The projection Problem* (4.7) *admits a closed-form solution given below:*

$$w_K^j = \begin{cases} S_K^j, \text{if } j \in \Omega_K \\ 0, \text{otherwise} \end{cases} \tag{4.8}$$

and

$$w_D^j = \begin{cases} S_D^j, \text{if } j \in \Omega_D \\ 0, \text{otherwise} \end{cases} \tag{4.9}$$

where S_K consists of the first d rows of S, S_K^j is the j-th row of S_K, S_D consists of the last s rows of S, S_D^j is the j-th row of S_D, Ω_K is the index subset of $\{1, 2, \cdots, d\}$ of size u, including all rows of S_K that are among the top u rows of S_K in term of the length of the row vector, and Ω_D is the index subset of $\{1, 2, \cdots, s\}$ of size v, including all rows of S_D that are among the top v rows of S_D in term of the length of the row vector.

One remaining issue is how to estimate the step size, which determines the amount of movement made along a given search direction. In this work, the well-known Lipschitz criterion is applied to select the step size.

Table 4.1: Twitter datasets and gold standard report (GSR)

Country	**#Tweets (million)**	**News source**[1]	**#Events**
Brazil	57	O Globo; O Estado de So Paulo; Jornal do Brasil	451
Paraguay	8	ABC Color; Ultima Hora; La Nacon	563
Mexico	51	La Jornada; Reforma; Milenio	1217
Venezuela	45	El Universal; El Nacional; Ultimas Noticias	678

Table 4.2: Event forecasting performance comparison (Precision, Recall, F-measure)

method	Mexico	Venezuela	Paraguay	Brazil	All Countries
DQEF	0.56, 0.40, 0.47	0.57, 0.61, 0.59	0.90, 0.15, 0.26	0.37, 0.34, 0.35	0.54, 0.38, 0.45
LASSO-K	0.68, 0.32, 0.44	0.93, 0.18, 0.30	1.00, 0.17, 0.29	0.62, 0.44, 0.51	0.72, 0.28, 0.40
DQEF+LASSO	0.57, 0.49, 0.53	0.59, 0.64, **0.61**	1.00, 0.11, 0.20	0.42, 0.49, 0.45	0.55, 0.44, 0.49
LASSO	0.70, 0.36, 0.48	**0.94**, 0.19, 0.32	1.00, 0.17, 0.29	0.63, 0.43, 0.51	0.73, 0.30, 0.43
rMTFL-D	**0.96**, 0.12, 0.21	0.66, 0.42, 0.51	1.00, 0.02, 0.04	1.00, 0.07, 0.13	**0.77**, 0.15, 0.25
rMTFL-K	0.78, 0.45, 0.57	0.53, **0.68**, 0.60	0.93, 0.43, 0.59	**0.79**, 0.55, 0.65	0.71, 0.51, 0.59
rMTFL	0.70 0.70 0.70	0.54, 0.61, 0.57	**0.96**, 0.32, 0.48	0.71, 0.52, 0.60	0.68, 0.57, 0.62
CMTFL-I	0.59, 0.87, 0.70	0.51, 0.66, 0.58	0.95, 0.39, 0.55	0.72, 0.60, **0.66**	0.62, 0.68, 0.65
CMTFL-II	0.71, 0.79, 0.75	0.53, 0.57, 0.55	0.78, **0.81**, **0.79**	0.76, 0.57, 0.65	0.69, **0.71**, **0.70**
CMTFL-III	0.71, **0.82**, **0.76**	0.53, **0.68**, **0.61**	0.85, 0.55, 0.67	0.53, **0.71**, 0.61	0.65, **0.71**, 0.68

4.6 Experiments

In this section, the performance of the three multi-task learning formulations is evaluated. First, the effectiveness and efficiency of the methods on real data are examined in comparison with baseline methods on multiple event forecasting tasks. Then, the parameter sensitivity of the methods is studied. Finally, several empirical case studies of civil unrest event forecasting are discussed to demonstrate the usefulness of these forecasting models. All the experiments were conducted on a 64-bit machine with Intel(R) core(TM) quad-core processor (i7CPU@3.40GHz) and 16.0GB memory.

4.6.1 Experiment Setup

The raw data was obtained by randomly sampling 10% (by volume) of the Twitter data from July 2012 to May 2013 in 4 countries in Latin America including Brazil, Paraguay, Mexico, and Venezuela, as shown in Table 4.1. Twitter data collection was partitioned into a sequence of date-interval subcollections. The Twitter data for the period from July 1,

[1]In addition to the top 3 domestic news outlets in each country, the following news outlets were included: The New York Times; The Guardian; The Wall Street Journal; The Washington Post; The International Herald Tribune; The Times of London; Infolatam.

Table 4.3: Run time comparison of different methods

	rMTFL	rMTFL-D	rMTFL-K	DQEF	LASSO-K	DQEF+LASSO	LASSO	CMTFL-I	CMTFL-II
Train (sec)	10.73	8.79	10.60	2.30	6.53	6.56	6.96	8.85	8.064
Test (sec)	0.003	0.001	0.003	0.01	0.001	0.001	0.001	0.003	0.01

2012 to December 31, 2012 was used for training while the second half of the period, from January 1, 2013 to May 31, 2013, was used for the performance evaluation. The locations of the tweets were geocoded by the geocoder in [86]. The event forecasting results were validated against a labeled events set, called the gold standard report (GSR), which was exclusively provided by MITRE [74]. GSR is a collection of civil unrest news reports from the most influential newspapers outlets in Latin America [109], as shown in Table 4.1. An example of a labeled GSR event is given by the tuple: (CITY="Hermosillo", STATE = "Sonora", COUNTRY = "Mexico", DATE = "2013-01-20").

In this experiment, two types of features were utilized. As introduced above, the first type of features is static features, which examines the relevance of tweets to fixed keywords. Specifically, they are defined as the daily counts of the keywords in the tweets. These keywords include 614 civil unrest related words (such as "protest" and "riot"), 192 phrases (such as "election fraud"), and country-specific actors (e.g., political parties and public figures). For each keyword, its translations in Spanish, Portuguese, and English are all included. The second type is dynamic features, which examines the volume of tweets containing dynamic keywords. Specifically, dynamic features are a set of counts, where each count is the number of daily tweets containing any of the top k ($k \in [1, s]$) dynamic keywords. The dynamic keywords were extracted and ranked based on dynamic query expansion (DQE) [109], which utilizes both semantic and social relationship to expand real-time keywords from seed query, as introduced in Section 4.3. The seed query includes: "protest", "march", "movement", "patriotic", "manifest", and their translations in Spanish and Portuguese. In this experiment, s was set to 20. Thus we have 20 dynamic features.

In the experiment, given the day-by-day tweets data, the event forecasting task is to predict whether there is an event or not in the next day for a specific city. To perform this task, a training set and a testing set were created for each city, where each data sample is the daily tweet observation with the above-mentioned features. On the training set, the label for each data sample was set as "1" if there is event on next day; and "0", otherwise. Three standard performance metrics are used for comparison: precision, recall, and F-measure. The predicted events were structured as tuples of (date, city). A predicted event is matched to a GSR event if both the date and city attributes are matched; Otherwise, it is considered as a false forecasting.

The following methods are included for performance comparison: 1). **LASSO** [94]. For each city, three LASSO models are trained utilizing different sets of features: i). both static and dynamic features, and ii). Only static features (denoted as **LASSO-K**). The regularization parameters of these models for different cities are set based on 10-fold cross validation. 2).

DQE-based event forecasting (**DQEF**). This model only considers the dynamic features, as introduced in Section 4.3. The number of top dynamic keywords, k, and the tweet count threshold γ are set for each countries by 10-fold cross-validation on training set. 3). **DQEF+LASSO**. For each city, it first uses DQEF method to do forecasting. If there is no predicted event, i.e., $Y_{l,t} = 0$, the LASSO model using only static features will be employed for forecasting. 4). Regularized Multi-task Feature Learning Model (**rMTFL**). For each country, a rMTFL model is built where each task is the event forecasting for a city. This model utilizes three sets of features: i). Both static and dynamic features, ii). Only static features (denoted as **rMTFL-K**);, and iii). Only dynamic features (denoted as **rMTFL-D**). The regularization parameters ρ_1 and ρ_{L2} are set based on 10-fold cross-validation. 5). Constrained multi-task feature learning model I (**CMTFL-I**). For each country, a model is built where each task is the event forecasting for a city. All the tasks share the same features, i.e., both the static and dynamic features. The feature number constraint r and the regularization parameter ρ_1 are set based on 10-fold cross-validation. 6) Constrained multi-task feature learning model II (**CMTFL-II**). For each country, a model is built where each task is the event forecasting for a city. All the tasks share the same features, i.e., the static and dynamic features. The 10-fold cross-validation was used to set the regularization parameter ρ_1, the numbers of static features u, and dynamic features v for each country. The sensitivity of these three parameters are studied in Section 4.6.3.

4.6.2 Performance

Table 4.2 summarizes the comparison among the proposed methods and the competing methods for the task of civil unrest event forecasting. These results showed that the methods that utilize both sets of static features and dynamic features performed better than the ones utilizing either one of them. For example, rMTFL outperformed rMTFL-D and rMTFL-K by 50% and 10% in F-measure, respectively. DQEF+LASSO and LASSO outperformed LASSO-K by 10% on average in F-measure. All these results demonstrated effectiveness of combining both type of features for event forecasting. Among all methods, CMTFL-II achieved a recall of 0.71 and a F-measure of 0.70, which were both the best. Moreover, the proposed CMTFL-II performed well consistently across different countries by being the best in Mexico and Paraguay, and competitive in Venezuela and Brazil. Other methods like the proposed CMTFL-I and rMTFL also obtained high F-measures, around 0.65, but not as competitive as the CMTFL-II. The reason is because (1) CMTFL-II is able to ensure the inclusion of both type of features, whose combination are demonstrated to be more effective than using either one of them, and (2) unlike rMTFL and CMTFL-I, CMTFL-II treats both types of features separately in the constraint based on their different characteristics, leading to a more effective integration of these two types of features. Finally, we can observe from Table 4.2 that the multi-task models outperformed the traditional LASSO models by 50% on average. This revealed the advantage of multi-task models, which can select features by learning from similar forecasting tasks for all the cities. The generalization and stability of

Table 4.4: Top 10 static features (translated in English) and the selection of dynamic features. TRUE means there is at least one dynamic feature selected; FALSE means no dynamic feature selected. rMTFL and CMTFL-II can ensure sufficient and stable selection of static features. CMTFL-II can ensure the selection of effective dynamic feature(s).

		Mexico					Brazil		
Methods	Features	Mexico City	Cuernavaca	Guadalajara	Morelia	Oaxaca	Bras lia	Rio de Janeiro	So Paulo
rMTFL	Static	fight	fight	remember	employ	university	participant	expensive	prisoners
		movement	hate	street	remember	allow	increased	strength	expensive
		election	hungry	work	unit	work	expensive	gringo	increase
		president	street	hate	water	develop	prepare	cries	cries
		congress	sent	president	university	hatred	include	progress	force
		initiative	calling	unit	change	problem	protest	participant	include
		progress	hungry	poor	class	progress	strength	protest	censorship
		hard	work	permit	statement	released	march	student	progress
		help	eliminate	killing	force	congress	gringo	censorship	prepare
		government	forcibly	remove	problem	killing	screams	include	student
	Dynamic	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
LASSO	Static	block	complaint	request	request	help	send	problem	throw
		fight	gunfire	confront	meet	power	power	water	bond
		work	tranquility	water	water	avoid	food	official	unit
		help	forward	danger	danger		forward	work	defeat
		hearsay	power	results	results		money	fight	send
		president	avoid	order	order		street	government	forward
		initiation		help	help			national	control
		occupy		national	national			employ	confront
		request			initiation				expensive
		power			town				finish
	Dynamic	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE
CMTFL-II	Static	protest	police	university	movement	block	shooting	attack	march
		fight	protest	expected	occupy	money	order	block	resolve
		president	struggle	movement	encounter	encounter	movement	occupy	attack
		government	patriot	manifest	hunger	memories	throw	arrest	warrant
		movement	movement	occupy	national	change	government	control	payment
		death	hunger	hate	change	police	submit	kill	poor
		poor	student	change	request	occupy	march	followers	claim
		national	block	class	fear	steal	national	throw	block
		expected	work	block	money	fight	block	ask	hatred
		wait	memories	official	country	president	attack	march	problem
	Dynamic	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

the forecasting performance can be improved by learning models for different cities together, especially for those cities that lack sufficient training samples.

Table 4.3 shows the run time of all the methods in training and testing. The training time of the multi-task models is only slightly larger than that of the LASSO model. As expected, the models using both static and dynamic features tend to consume more time than the ones only using either type of features. All methods consumed negligible testing time (around 0.01 0.003 sec).

Table 4.4 shows the specific features selected by different models, including rMTFL, LASSO, and the proposed CMTFL-II for several cities of two countries, i.e., Mexico (Spanish-spoken) and Brazil (Portuguese-spoken). According to Table 4.4, CMTFL-II effectively selected static features (i.e., keywords) very relevant to civil unrest, and the selection was stable and consistent across different cities. Moreover, the selection of dynamic feature(s) was ensured,

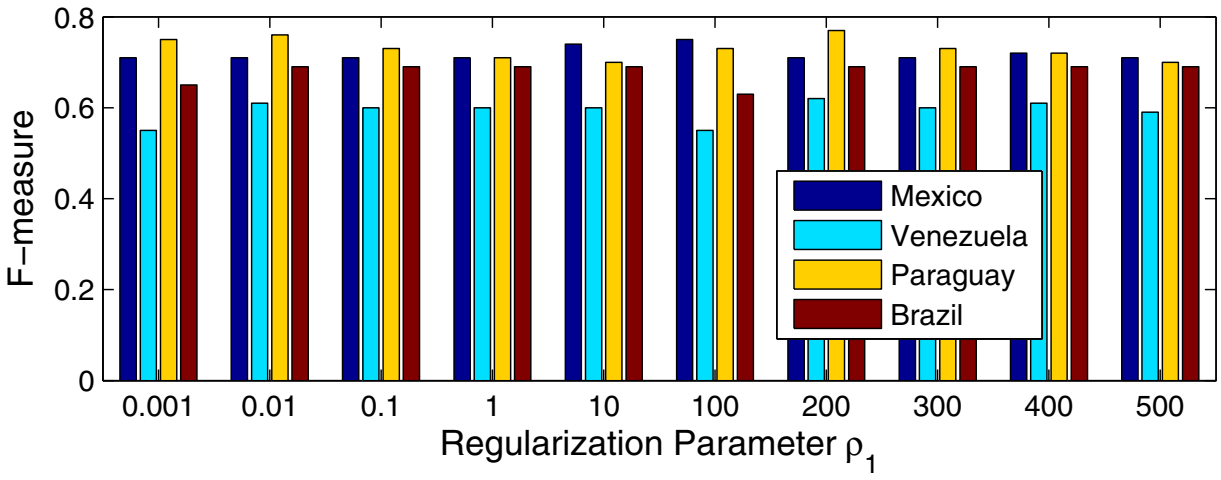


Figure 4.3: Sensitivity analysis on the regularization parameter.

as shown in the bottom row, which enhanced the capacity to consider the burstiness of tweets containing dynamic keywords. rMTFL model also effectively selected civil unrest-related keywords as the top static features. However, it cannot guarantee the selection of dynamic features because in all the listed cities for Brazil, it did not select any dynamic features. The selected static features for LASSO model was not consistent across different cities, and more importantly, not as relevant and sufficient as the above-two multi-task learning models in several cities, especially the smaller ones, such as Oaxaca and Cuernavaca. Additionally, the selection of dynamic features was not ensured, such as in Morelia and Braslia.

4.6.3 Parameter Sensitivity Study

There are three main parameters in the proposed rMTFL II model, which are the regularization parameter ρ_1, number of selected static features u, and number of selected dynamic features v.

Figure 4.3 illustrates the performance of the proposed model versus, ρ_1, the regularization parameter. By varying ρ_1 in a large range from 0.001 to 500, the performance in F-measures for all the 4 countries are stable. The fluctuation ranges are typically within 8%.

Figure 4.4 shows the sensitivity results of varying u, the number of selected static features from 10 to 100. In general, for all the countries, the F-measures at $u = 10$ and $u = 20$ are slightly lower than other cases, but after u is larger than 30, the F-measure becomes stable. This is because a small number of selected static features may not capture the complexity of the event forecasting task. Thus, the number of selected keywords should not be too small. But when the selected static features are sufficient (¿30), using more of them does not necessarily lead to additional performance improvement.

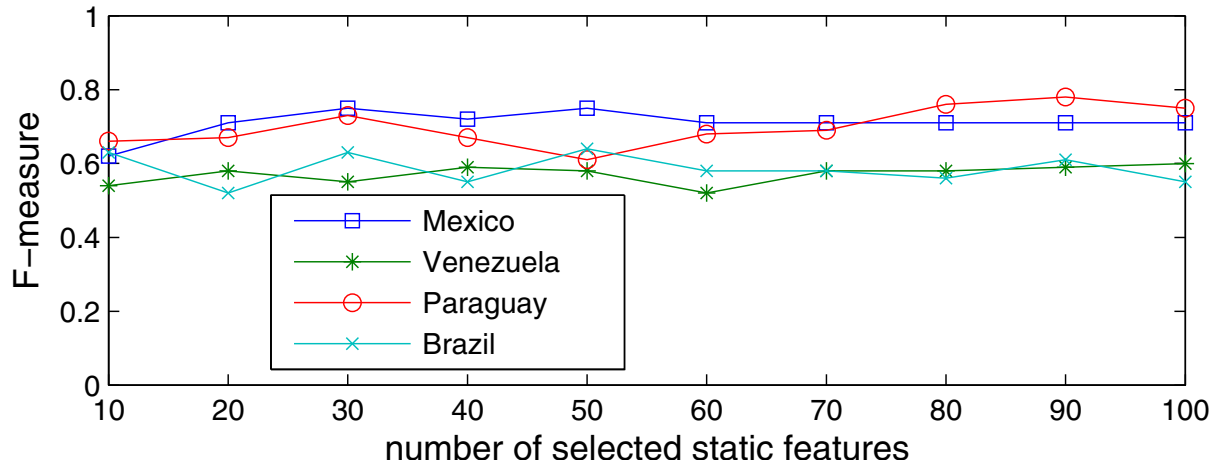


Figure 4.4: Sensitivity analysis on the number of selected static features.

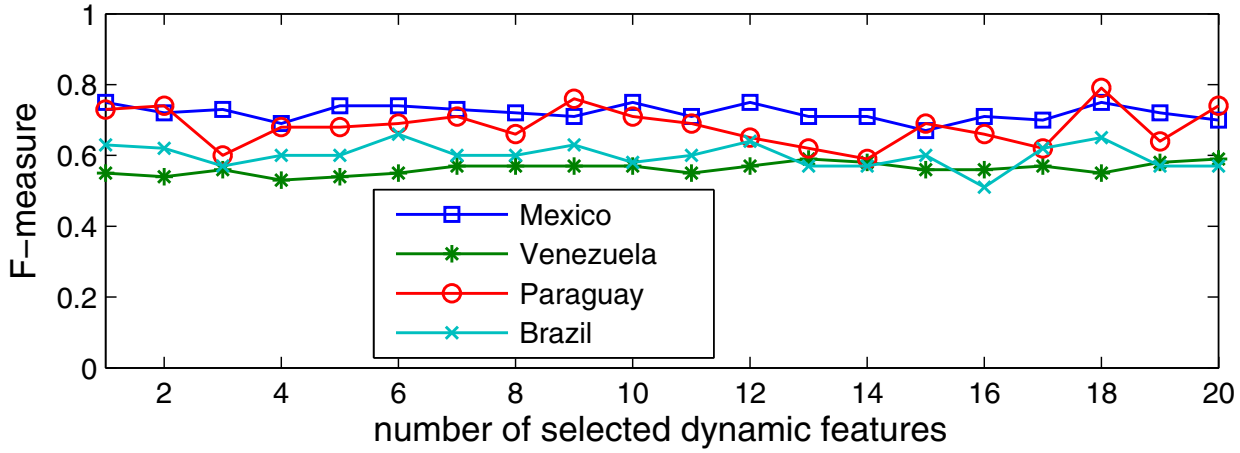


Figure 4.5: Sensitivity analysis on the number of selected dynamic features.

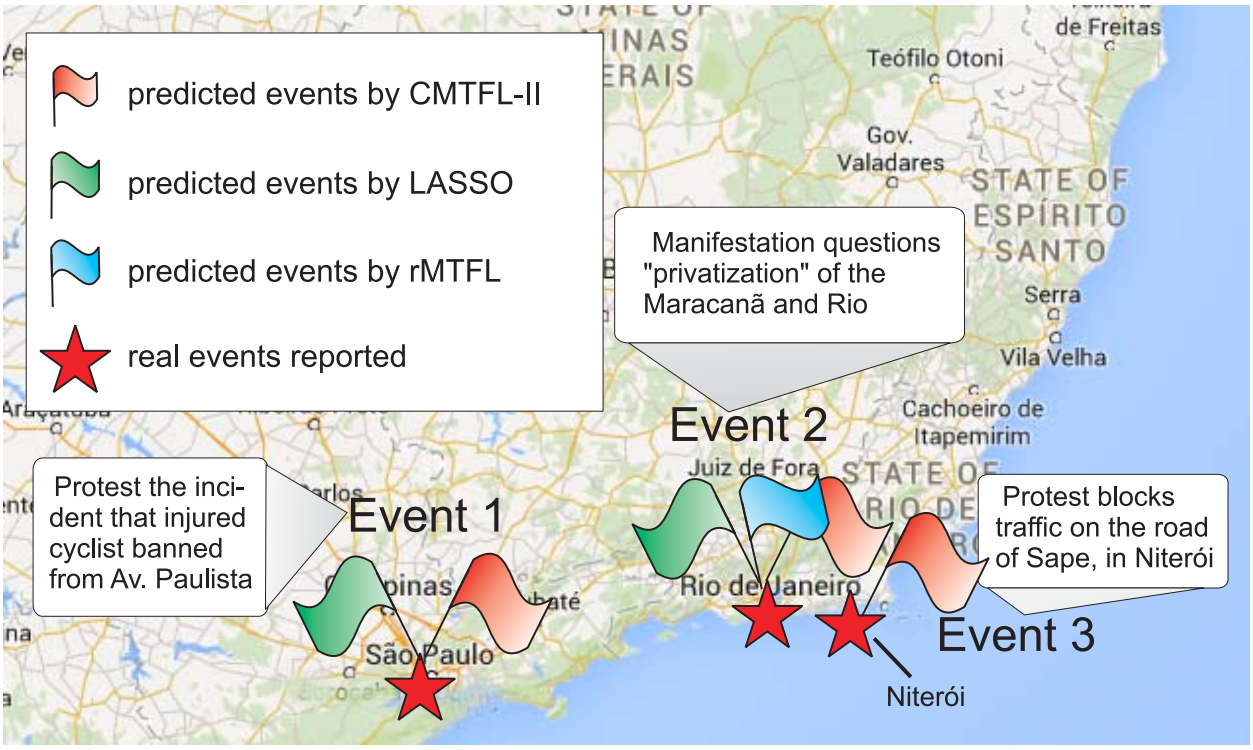


Figure 4.6: A map of civil unrest events and forecasting hotspots on March 17th, 2013 in Brazil.

Figure 4.5 illustrates the F-measures obtained by varying v, the number of selected dynamic features, from 1 to 20. The F-measure is quite stable, even when v is as small as 1, which demonstrates that even only 1 dynamic feature could be sufficient to capture the dynamic in the civil unrest tweets, and adding more dynamic features does not add extra information.

4.6.4 Case Studies

Numerous interesting events predicted by the proposed approaches, CMTFL-I and CMTFL-II were observed in the experiments. For instance, Figures 4.6 and 4.7 record two waves of civil unrest events that occurred on March 17th, 2013 in Brazil, and April 17th, 2013 in Paraguay, respectively.

We can observe from Figure 4.6 that there were three events in Brazil, among which Event 1 and Event 2 happened in large cities, e.g., Sao Paulo and Rio de Janeiro, while Event 3 was in a smaller city, Niteri. Note that the city Niteri does not have any training sample. The proposed CMTFL-II successfully predicted all of events, even for the city Niteri. This is because CMTFL-II jointly learned the models of all the tasks (i.e., cities). Even the model of the city has no training sample, it can still be estimated by data from other cities. The LASSO model predicted two of them but failed on the forecasting of Event 3. This is

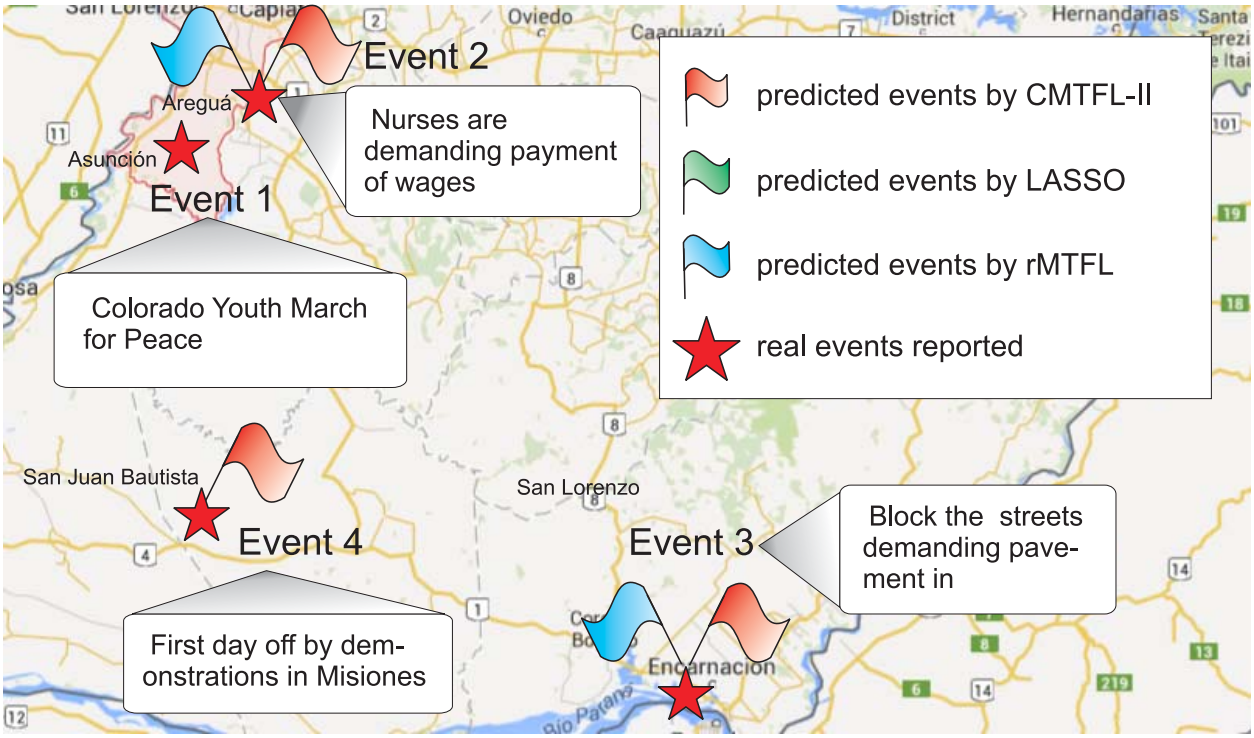


Figure 4.7: A map of civil unrest events and forecasting hotspots on April 17, 2013 in Paraguay.

because each LASSO model is trained for each city individually, and thus the events of the city without any training sample cannot be predicted. The rMTFL model only predicted one event for Rio De Janerio. Its failure of discovering events for two other cities might be due to its exclusion of the dynamic features after training, as shown in Table 4.4. This reduces its capability to uncover the burstiness of dynamic keywords. This further verifies the need for a separate selection of the static and dynamic features as in the proposed CMTFL-II model.

We can observe from Figure 4.7 that there were four events in Paraguay, among which Event 2, Event 3, and Event 4 had been successfully predicted by CMTFL-II. rMTFL predicted Event 2 and Event 3 while LASSO failed to predict any event. As shown in Table 4.1, Paraguay is a country that the number of reported events is large but the volume of tweets is relatively small, i.e., the ratio of #tweets/#events is less than one third of other countries. The sparsity of tweets data make the forecasting more difficult for Paraguay by methods without using multi-task learning, as shown in Table 4.2.

4.7 Conclusions

This work presents a novel multi-task learning framework to the problem of spatial event forecasting in Social Media. Existing methods are not able to concurrently address the critical challenges, such as dynamic patterns of features, and geographic heterogeneity. The proposed work considers the estimation of predictive models in different locations as a multi-task learning problem, in order to use the shared information between locations, which effectively increases the sample size for each location. This work further models both static and dynamic features using different constraints to balance both homogeneity and diversity between these two types of features. Efficient algorithms are proposed based on the IGHT that are able to run in real time. The empirical results demonstrated that we can effectively detect civil unrest events, outperforming competing methods by a substantial margin on both precision and recall. For the future work, the proposed multi-task learning framework are expected to be extended by exploring more complex relationships between locations and integrating human domain knowledge as priors.

Chapter 5

Multi-source Feature Learning for Spatial Event Forecasting

Multi-source event forecasting has proven promising but still suffers from several challenges, including 1) geographical hierarchies in multi-source data features, 2) missing values, and 3) characterization of structured feature sparsity. This chapter proposes a novel feature learning model that concurrently addresses all the above challenges. The research background is introduced in Section 5.1. Section 5.2 reviews background and related work, and Section 5.3 introduces the problem setup. Section 5.4 presents the HIML model and an efficient model parameter optimization algorithm. The experiments on 10 real-world datasets are presented in Section 5.5, and the paper concludes with a summary of the research in Section 5.6.

5.1 Introduction

Significant societal events such as disease outbreaks and mass protests have a tremendous impact on the entire society, which strongly motivates anticipating their occurrences in advance. For example, according to a recent World Health Organization (WHO) report [26], seasonal influenza alone is estimated to result in around 4 million cases of severe illness and about 250,000 to 500,000 deaths each year. In regions such as the Middle East and Latin America, the majority of instabilities arise from extremism or terrorism, while others are the result of civil unrest. Population-level uprisings by disenchanted citizens are generally involved, usually resulting in major social problems that may involve economic losses that run into the billions of dollars and create millions of unemployed people. Significant societal events are typically caused by multiple social factors. For example, civil unrest events could be caused by economic factors (e.g., increasing unemployment), political factors (e.g., a presidential election), and educational factors (e.g., educational reform). Moreover, societal events can also be driven and orchestrated through social media and news reports. For

example, in a large wave of mass protests in the summer of 2013, Brazilian protesters calling for demonstrations frequently used Twitter as a means of communication and coordination. Therefore, to fully characterize these complex societal events, recent studies have begun to focus on utilizing indicators from multiple data sources to track different social factors and public sentiment that jointly indicate or anticipate the potential future events.

These multi-source based methods share essentially similar workflows. They begin with collecting and preprocessing each single data source individually, from which they extract meaningful features such as ratios, counts, and keywords. They then aggregate these feature sets from all different sources to generate the final input of the forecasting model. The model response, in this case predicting the occurrence of future events, is then mapped to these multi-source input features by the model. Different data sources commonly have different time ranges. For example, Twitter has been available since 2006, but CDC data dates back to the 1990s. When the predictive model utilizes multiple data sources, of which some are incomplete, typically the samples with missing values in any of these data sources are simply removed, resulting in substantial information loss.

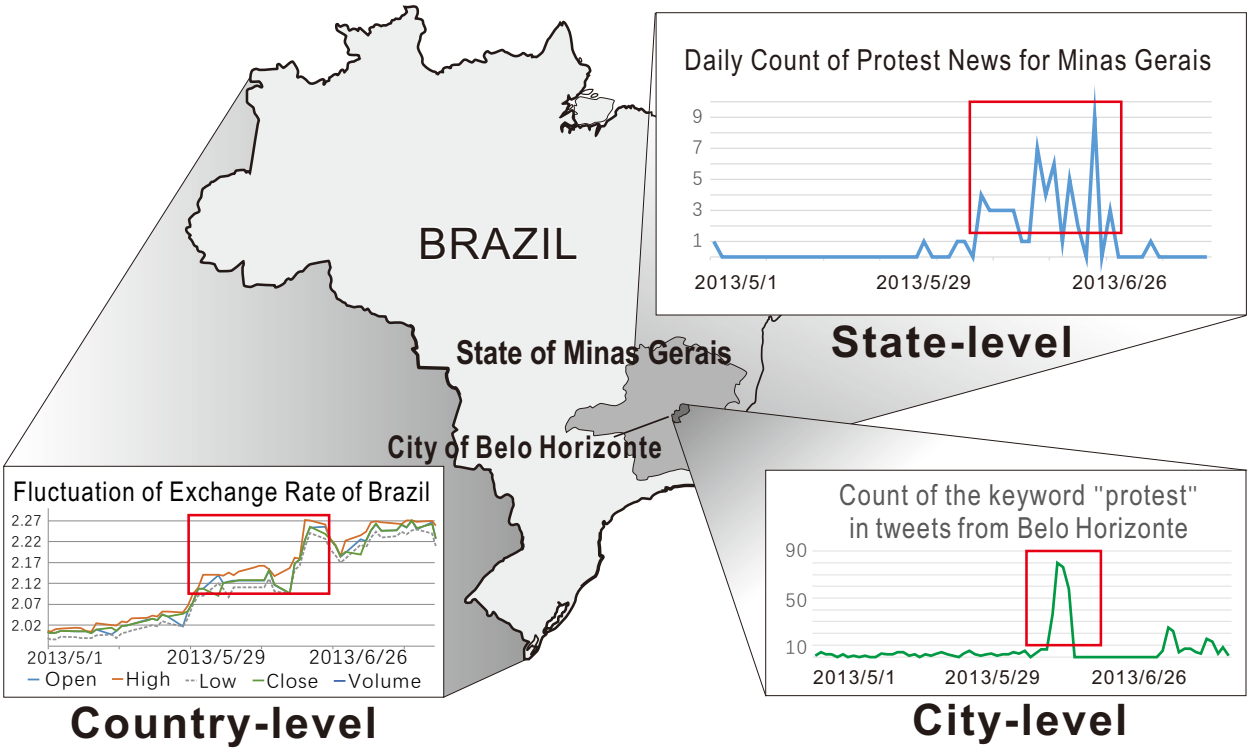


Figure 5.1: Predictive indicators from multiple data sources with different geographical levels during the "Brazilian Spring" civil unrest movement.

Multi-source forecasting of significant societal events is thus a complex problem that currently still faces several important challenges. **1. Hierarchical topology**. When features in different data sources come from different topological levels, they cannot normally be treated as independent and homogeneous. For example, Figure 5.1 shows multiple indica-

tors during the "Brazilian Spring", the name given to a large wave of protest movements in Brazil in June 2013 caused by economic problems and spread by social media. Here, indicators in economy and social media would be the precursors of the protests. Some of these indicators are country-level, such as the exchange rate; some are state-level, such as news reports specific to a state; and some are city-level, such as the Twitter keyword count for chatter geolocated to a specific city. When forecasting city-level protest events, however, it is unrealistic to simply treat the union of all these multi-level features directly as city-level features for prediction. Moreover, it is unreasonable to assume that all cities across the country are equally influenced by the higher level features and are completely independent of each other. **2. Interactions involving missing values**. When features are drawn from different hierarchical topologies, features from higher levels influences those lower down. Thus, the missing value in such feature sets will also influence other features. This means that simply discarding the missing values is not an ideal strategy as its interactions with other features also need to be considered. **3. Geo-hierarchical feature sparsity**. Among the huge number of features from multiple data sources, only a portion of them will actually be helpful for predicting the response. However, due to the existence of hierarchical topology among the features, as mentioned earlier, features are not independent of each other. It is thus clearly beneficial to discover and utilize this hierarchically structured pattern to regulate the feature selection process.

In order to simultaneously address all these technical challenges, this chapter presents a novel model named hierarchical incomplete multi-source feature learning (HIML). HIML is capable of handling the features' hierarchical correlation pattern and secure the model's robustness against missing values and their interactions. To characterize the hierarchical topology among the features from multi-source data, a multi-level model is built that can not only handle all the features' impacts on the response, but also take into account the interactions between higher- and lower-level features. Under the assumption of feature sparsity, we characterize the hierarchical structure among the features and utilize it to regulate a proper hierarchical pattern. The HIML model can also handle missing values among multiple data sources by incorporating a multitask strategy that treats each missing pattern as a task.

The main contributions of this study are summarized below.

- **Design a framework for event forecasting based on hierarchical multi-source indicators**. A generic framework is proposed for spatial event forecasting that utilizes hierarchically topological multiple data sources and is based on a generalized multi-level model. A number of classic approaches on related research are shown to be special cases of the proposed model.
- **Propose a robust model for geo-hierarchical feature selection**. To model the structured inherent in geo-hierarchical features across multiple data sources, this work proposes an N-level interactive group Lasso based on strong hierarchy. To handle interactions among missing values, the proposed model adopts a multitask framework that

is capable of learning the shared information among the tasks corresponding to all the missing patterns.

- **Develop an efficient algorithm for model parameter optimization.** To learn the proposed model, a constrained overlapping group lasso problem needs to be solved, which is technically challenging. By developing an algorithm based on the alternating direction method of multipliers (ADMM) and introducing auxiliary variables, a globally optimal solution to this problem is ensured.
- **Conduct extensive experiments for performance evaluations.** The proposed method was evaluated on 10 different datasets in two domains: forecasting civil unrest in Latin America and influenza outbreaks in the United States. The results demonstrate that the proposed approach runs efficiently and consistently outperforms the best of the existing methods along multiple metrics.

5.2 Related Work

This section introduces related work in several research areas.

Event detection and forecasting in social media. There is a large body of work that focuses specifically on the identification of ongoing events, such as earthquakes [90] and disease outbreaks [111]. Unlike these approaches, which typically uncover events only after their occurrence, event forecasting methods predict the incidence of such events in the future. Most event forecasting methods focus on temporal events, with no interest in the geographical dimension, such as elections [78] and stock market movements [21]. Few existing approaches can provide true spatiotemporal resolution for the predicted events [109]. For example, Gerber utilized a logistic regression model for spatiotemporal event forecasting [41]. Zhao et al. [112] designed a multitask learning framework that models forecasting tasks in related geo-locations concurrently. Zhao et al. [110] also designed a new predictive model that jointly characterizes the temporal evolution of both the semantics and geographical burstiness of social media content.

Multi-source event forecasting. In recent years, a few researchers have begun to utilize multiple data sources as surrogates to forecast future significant societal events such as disease outbreaks and civil unrest. Chakraborty et al. proposed an ensemble model to forecast Influenza-like Illness (ILI) ratios based on seven different data sources [30]. Focusing on civil unrest events, Ramakrishnan et al. employ a LASSO model as the event predictor, where the inputs are the union of feature sets from different data sources [86]. Kallus explores the predictive power of news, blogs, and social media for political event forecasting [57]. However, although these models utilize multiple data sources that can be used to indicate a number of different aspects of future events, they typically ignore the potential relationships, topology, and hierarchy among these multi-source features.

Missing values in multiple data sources. The prevention and management of missing data has been discussed and investigated in earlier work [45]. One category of work focuses on estimating missing entries based on the observed values [40]. These methods work well when missing data are rare, but are less effective when a significant amount of data is missing. To address this problem, Hernandez et al. utilized probabilistic matrix factorization [50], but their method is restricted to non-random missing values. Yuan et al. [108] utilized multitask learning to learn a consistent feature selection pattern across different missing groups. However, none of these approaches focus specifically on missing values in hierarchical multiple data sources.

Feature selection in the presence of interactions. Feature selection by considering feature interactions has been attracting research interest for some time. For example, to enforce specific interaction patterns, Peixoto et al. [47] employed conventional step-wise model selection techniques with hierarchical constraints. Unfortunately such approaches are expensive for high-dimensional data. Choi et al. proposed a more efficient LASSO-based non-convex problem with re-parametrized coefficients [34]. To obtain globally optimal solutions, more recent research has utilized interaction patterns such as strong or weak hierarchy that are enforced via convex penalties or constraints. Both of these apply a group-lasso-based framework; Lim and Hastie [66] work with a combination of continuous and categorical variables, while Haris et al. [46] explore different types of norms. However, none of these approaches considers missing values in the feature sets.

5.3 Problem Setup

In this section, the problem addressed by this research is formulated. Specifically, Section 5.3.1 poses the hierarchical multi-source event forecasting problem and introduces the multi-level model formulation. Section 5.3.2 discusses the problem generalization and challenges.

5.3.1 Problem Formulation

Multiple data sources could originate at different geographical levels, for example city-level, state-level, or country-level, as shown in Figure 5.1. Before formally stating the problem, this section first introduces two definitions related to geographical hierarchy.

Definition 3 (Subregion)**.** *Given two locations q_i and s_j under the ith and jth $(i < j)$ geographical levels, respectively, if the whole spatial area of the location q_i is included by location s_j, we say q_i is a* ***subregion*** *of s_j, denoted as $q_i \sqsubseteq s_j$ or equally $s_j \sqsupseteq q_i$ $(i < j)$.*

Definition 4 (Location Tuple)**.** *The location of a tweet or an event is denoted by a* ***location tuple*** *$l = (l_1, l_2, \cdots, l_N)$, which is an array that configures each location l_n in each geo-level n in terms of a parent-child hierarchy such that $l_{n-1} \sqsubseteq l_n (n = 2, \cdots, N)$, where l_n is the* ***parent*** *of l_{n-1} and l_{n-1} is the* ***child*** *of l_n.*

For example, for the location "San Francisco", its location tuple could be ("San Francisco", "California", "USA") that consists of this city, its parent, and the parent's parent.

Suppose X denotes the set of multiple data sources coming from N different geographical levels. These can be temporally split into fixed time intervals t (e.g., "date") and denoted as $X = \{X_{t,l}\}_{t,l}^{T,L} = \{X_{t,l_n}\}_{t,l,n}^{T,L,N}$, where $X_{t,l_n} \in \mathbb{N}^{|\mathcal{F}_n|\times 1}$ refers to the feature vector for the data at time t in location l_n under nth geo-level. Specifically, the element $[X_{t,l_n}]_i$ $(i \neq 0)$ is the value for ith feature while $[X_{t,l_n}]_0 = 1$ is a dummy feature to provide a compact notation for bias parameter in forecasting model. T denotes all the time intervals. L denotes the set of all the locations and N denotes the set of all the geographical levels. $\mathcal{F}_n$ denotes the feature set for Level n and $\mathcal{F} = \{\mathcal{F}_n\}_{n=1}^{N}$ denotes the set of features in all the geo-levels. We also utilize a binary variable $Y_{t,l} \in \{1, 0\}$ for each location $l = (l_1, \cdots, l_N)$ at time t to indicate the occurrence ('yes' or 'no') of a future event. We also define $Y = \{Y_{t,l}\}_{t,l}^{T,L}$. Thus, the hierarchical multi-source event forecasting problem can be formulated as below:

Problem Formulation: For a specific location $l = (l_1, \cdots, l_N)$ at time t, given data sources under N geographical levels $\{X_{t,l_1}, \cdots, X_{t,l_N}\}$, the goal is to predict the occurrence of future event $Y_{\tau,l}$ where $\tau = t + p$ and $p > 0$ is the lead time for forecasting. Thus, the problem is formulated as the following mapping function:

$$f : \{X_{t,l_1}, \cdots, X_{t,l_N}\} \rightarrow Y_{\tau,l} \tag{5.1}$$

where f is the forecasting model.

In Problem (6.1), input variables $\{X_{t,l_1}, \cdots, X_{t,l_N}\}$ are not independent of each other because the geographical hierarchy among them encompasses hierarchical dependence. Thus classical single-level models such as linear regression and logistic regression cannot be utilized here.

As generalizations of the single-level models, multi-level models are commonly used for problems where input variables are organized at more than one level. The variables for the locations in Level n–1 are dependent on those of their *parents*, which are in Level n $(2 \leq n \leq N)$. The highest level (i.e., Level N) variables are independent variables. Without loss of generality and for convenience, here we first formulate the model with $N = 3$ geographical levels (e.g., city-level, state-level, and country-level) and then generalize it to $N \in Z^+$ in Section 5.3.2. The multi-level models for hierarchical multi-source event forecasting are formulated as follows:

$$\begin{aligned}
&(level-1) \quad Y_{\tau,l} = \alpha_0 + \sum\nolimits_{i=1}^{|\mathcal{F}_1|} \alpha_i^T \cdot [X_{t,l_1}]_i + \varepsilon \\
&(level-2) \quad \alpha_i = \beta_{i,0} + \sum\nolimits_{j=1}^{|\mathcal{F}_2|} \beta_{i,j}^T \cdot [X_{t,l_2}]_j + \varepsilon_i \\
&(level-3) \quad \beta_{i,j} = W_{i,j,0} + \sum\nolimits_{k=1}^{|\mathcal{F}_3|} W_{i,j,k}^T \cdot [X_{t,l_3}]_k + \varepsilon_{i,j}
\end{aligned} \tag{5.2}$$

where α_i, $\beta_{i,j}$, and $W_{i,j,k}$ are the coefficients for models of Level 1, Level 2, and Level 3, respectively. Each Level-1 parameter α_i is linearly dependent on Level-2 parameters $\beta_{i,j}$ and each Level-2 parameter $\beta_{i,j}$ is again linearly dependent on Level-3 parameters $W_{i,j,k}$. ε, ε_i,

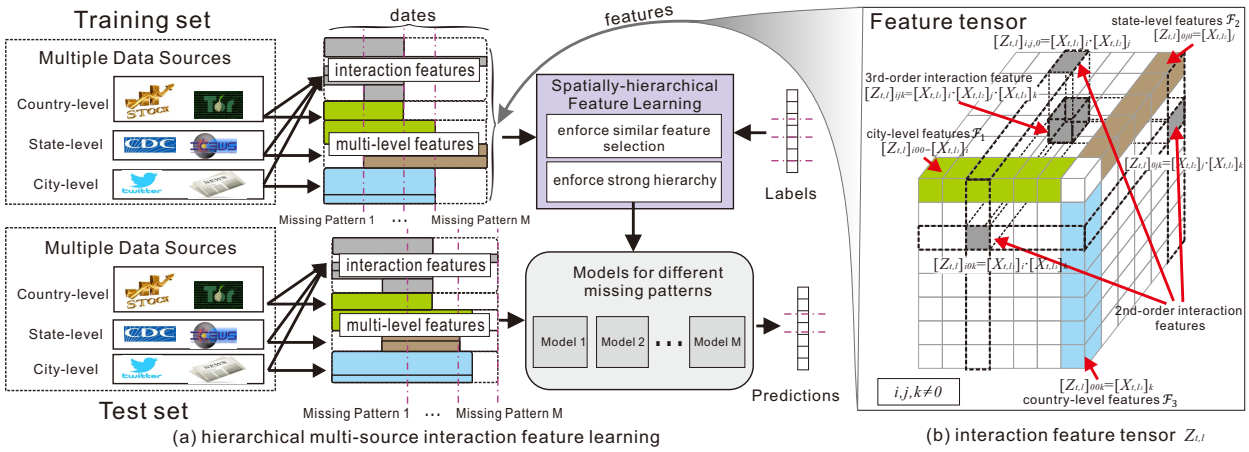


Figure 5.2: A schematic view of hierarchical incomplete multi-source feature learning (HIML) model.

and $\varepsilon_{i,j}$ are the noise terms for Levels 1, 2, and 3. Combining all the formulas in Equation (5.2), we get:

$$Y_{\tau,l} = \sum_{i=0}^{|\mathcal{F}_1|}\sum_{j=0}^{|\mathcal{F}_2|}\sum_{k=0}^{|\mathcal{F}_3|} W_{i,j,k}\cdot[X_{t,l_1}]_i\cdot[X_{t,l_2}]_j\cdot[X_{t,l_3}]_k + \varepsilon \tag{5.3}$$

where ε is noise term. Utilizing tensor multiplication, Equation (5.3) can be expressed in the following compact notation:

$$Y_{\tau,l} = W \odot Z_{t,l} + \varepsilon \tag{5.4}$$

where $W = \{W_{i,j,k}\}_{i,j,k=0}^{|\mathcal{F}_1|,|\mathcal{F}_2|,|\mathcal{F}_3|}$ and $Z_{t,l}$ are two $(|\mathcal{F}_1|+1)\times(|\mathcal{F}_2|+1)\times(|\mathcal{F}_3|+1)$ tensors, and an element of $Z_{t,l}$ is defined as $[Z_{t,l}]_{i,j,k} = [X_{t,l_1}]_i\cdot[X_{t,l_2}]_j\cdot[X_{t,l_3}]_k$. The operator $\odot$ is the summation of the Hadamard product of two tensors such that $A\odot B = \sum_{i,j,k} A_{ijk}\cdot B_{ijk}$ for 3rd-order tensors A and B.

The tensor $Z_{t,l}$ is illustrated in Figure 5.2(b). Specifically, the terms $[Z_{t,l}]_{i,0,0} = [X_{t,l_1}]_i$, $[Z_{t,l}]_{0,j,0} = [X_{t,l_2}]_j$, and $[Z_{t,l}]_{0,0,k} = [X_{t,l_3}]_k$ are the main-effect variables shown, respectively as green, blue, and brown nodes in Figure 5.2(b). Main-effect variables are independent variables. The terms $[Z_{t,l}]_{i,j,0} = [X_{t,l_1}]_i\cdot[X_{t,l_2}]_j$, $[Z_{t,l}]_{i,0,k} = [X_{t,l_1}]_i\cdot[X_{t,l_3}]_k$, and $[Z_{t,l}]_{0,j,k} = [X_{t,l_2}]_j\cdot[X_{t,l_3}]_k$ are 2nd-order interactive variables and are shown as nodes on the surfaces formed by the lines of the main-effect variables in Figure 5.2(b). Their values are dependent on both of their two main-effect variables. The terms $[Z_{t,l}]_{i,j,k} = [X_{t,l_1}]_i\cdot[X_{t,l_2}]_j\cdot[X_{t,l_3}]_k$ are called 3rd-order interactions because their values are dependent on 2nd-order interactive variables, as shown in Figure 5.2(b). Finally, denote $Z = \{Z_{t,l}\}_{t,l}^{T,L}$ as the set of feature tensors for all the locations L and time intervals T.

5.3.2 Problem Generalization

Here, the 3-level model in Equation (5.4) is generalized into an N-level model. Moreover, the linear function in Equation (5.4) is generalized into nonlinear setting.

1. N-level Geo-hierarchy

In Equation (5.4), it is assumed that the number of geographical levels is $N = 3$. Now this is extended by introducing the generalized formulation where the integer $N \geq 2$. The formulation in Equation (5.4) is retained, and the operator $\odot$ is generalized into a summation of the Nth-order Hadamard product such that $A \odot B = \sum_{i_1,\cdots,i_N} A_{i_1,\cdots,i_N} \cdot B_{i_1,\cdots,i_N}$. For simplicity, this can be denoted as $A \odot B = \sum_{\vec{i}} A_{\vec{i}} \cdot B_{\vec{i}}$, where $\vec{i} = \{i_1, i_2, \cdots, i_N\}$.

2. Generalized Multi-level Linear Regression

In Equation (5.4), it is assumed that a linear relation between input variable $Z_{t,l}$ and the response variable $Y_{t,l}$. However, in many situations, a more generalized relation could be necessary. For example, we may need a logistic regression setup when modeling a classification problem. Specifically, the generalized version of our multi-level model adds a nonlinear mapping between the input and response variables:

$$Y_{t,l} = h(W \odot Z_{t,l}) + \varepsilon \tag{5.5}$$

where $h(\cdot)$ is a convex and differentiable mapping function. In this chapter, the standard logistic function $h(x) = 1/(1 + e^{-x})$ is considered (see Section 5.4.3).

Although the models proposed in Equations (5.4) and (5.5) are capable of modeling the features coming from different geo-hierarchical levels, they suffer from three challenges: 1). The weight tensor W is typically highly sparse. This is because the main effects could be sparse, meaning that their interaction (i.e., multiplication) will be even more sparse. Without considering this sparsity, the computation will be considerably more time-consuming. 2). The pattern of W is structured. There is a geo-hierarchy among the multi-level features, which causes their interactions in W to follow specific sparsity patterns. A careful and effective consideration and utilization of this structure is both vital and beneficial. 3) The models do not consider missing values, whereas these are actually quite common in practical applications that use multi-source data. A model that is capable of handling missing values is therefore imperative. In the next section, HIML, a novel hierarchical feature learning approach based on constrained overlapping group lasso, is proposed to address all three challenges.

5.4 Hierarchical Incomplete Multi-source Feature Learning

Without loss of generality and for convenience, Section 5.4.1 first proposes the hierarchical feature learning model for $N = 3$ geographical levels, and then Section 5.4.2 generalizes it to handle the problem of missing values, as shown in Figure 5.2. Section 5.4.3 then takes the model further by generalizing it to $N \in \mathbb{Z}^+$ geographical levels and incorporating nonlinear loss functions. The algorithm for the model parameter optimization is proposed in Section 5.4.4. The relationship of the proposed HIML model to existing models is discussed in Section 5.4.5.

5.4.1 Hierarchical Feature Correlation

In fitting models with interactions among variables, a 2nd-order strong hierarchy is widely utilized [46, 56] as this can handle the interactions between two sets of main-effect variables. Here, their definition is introduced as follows:

Lemma 2 (2nd-order Strong Hierarchy)**.** *If a 2nd-order interaction term is included in the model, then both of its product factors (i.e., main effect variables) are present. For example, if* $W_{i,j,0} \neq 0$*, then* $W_{i,0,0} \neq 0$ *and* $W_{0,j,0} \neq 0$*.*

Here the 2nd-order Strong Hierarchy to Nth-order Strong Hierarchy ($N \in \mathbb{Z}^+ \wedge N \geq 2$) is generated as follows:

Theorem 2 (Nth-order Strong Hierarchy)**.** *If an* N*th-order interaction variable is included in the model, then all of its* n*th-order* ($2 \leq n < N$) *interactive variables and main-effect variables are included.*

Proof. According to Lemma 2, if an nth-order interaction variable ($2 \leq n \leq N$) is included, then its product-factor pairs, $(n-1)$th-order interaction factor and main effect, must also be included. Similarly, if an $(n-k)$th-order interaction variable ($1 \leq k \leq n-2$) is included, then so must its pairs of $(n-k-1)$th-order interaction factor and main effect. By varying k from 1 to $N-2$, it is immediately known that any nth-order ($2 \leq n < N$) interactive variables and main effects must be included. □

When $N = 3$, Theroem 2 becomes the *3rd-order strong hierarchy*. Specifically, if $W_{i,j,k} \neq 0$, then we have $W_{i,j,0} \neq 0$, $W_{i,0,k} \neq 0$, $W_{0,j,k} \neq 0$, $W_{i,0,0} \neq 0$, $W_{0,j,0} \neq 0$, and $W_{0,0,k} \neq 0$, where $i, j, k \neq 0$. In the following a general convex regularized feature learning approach is proposed to enforce the 3*rd-order strong hierarchy*.

The proposed feature learning model minimizes the following penalized empirical loss:

$$\min_{W} \mathcal{L}(W) + \Omega(W) \tag{5.6}$$

where $\mathcal{L}(W)$ is the loss function such that $\mathcal{L}(W) = \sum_{t,l} \|Y_{\tau,l} - W \odot Z_{t,l}\|_F^2$. $\Omega(W)$ is the regularization term that encodes task relatedness:

$$\begin{aligned}\Omega(W) =& \lambda_0 \sum_{i,j,k\neq 0} |W_{i,j,k}| + \lambda_1 \sum_{j+k\neq 0} \|W_{\cdot,j,k}\|_F \\ &+ \lambda_2 \sum_{i+k\neq 0} \|W_{i,\cdot,k}\|_F + \lambda_3 \sum_{i+j\neq 0} \|W_{i,j,\cdot}\|_F\end{aligned} \tag{5.7}$$

where $\|\cdot\|_F$ is the Frobenius norm. λ_0, λ_1, λ_2, and λ_3 are regularization parameters such that $\lambda_0 = \lambda/(|\mathcal{F}_1| \cdot |\mathcal{F}_2| \cdot |\mathcal{F}_3|)$, $\lambda_1 = \lambda/(\sqrt{|\mathcal{F}_1|} \cdot |\mathcal{F}_2| \cdot |\mathcal{F}_3|)$, $\lambda_2 = \lambda/(|\mathcal{F}_1| \cdot \sqrt{|\mathcal{F}_2|} \cdot |\mathcal{F}_3|)$, and $\lambda_3 = \lambda/(|\mathcal{F}_1| \cdot |\mathcal{F}_2| \cdot \sqrt{|\mathcal{F}_3|})$, where λ is a regularization parameter that balances the trade off between the loss function $\mathcal{L}(W)$ and the regularization terms. Equation (5.7) is a higher-order generalization of the ℓ_2 penalty proposed by Haris et al. [46], which enforces enforces a hierarchical structure under a 2nd-order strong hierarchy.

5.4.2 Missing Features Values in the Presence of Interactions

As shown in Figure 5.2(a), multiple data sources usually have different time durations, which result in incomplete data in multi-level features and about the feature interactions among them. Before formally describing the proposed generalized model for missing values, first, two related definitions are introduced.

Definition 5 (Missing Pattern Block)**.** *A **missing pattern block (MPB)** is a block of multi-source data $\{X_{t,l}\}_{t,l}^{T_m,L}$ ($T_m \subseteq T$) that share the same missing pattern of feature values. Define $\mathcal{M}(X_{t,l})$ as the set of missing-value features of the data $X_{t,l}$. Assume the total number of MPBs is M, then they must satisfy the following three criteria:*

- *(completeness)* : $T = \bigcup_m^M T_m$
- *(coherence)* : $\forall t_i, t_j \in T_m : \mathcal{M}(X_{t_i,l}) = \mathcal{M}(X_{t_j,l})$
- *(exclusiveness)* : $\forall t_i \in T_m, t_j \in T_n, m \neq n : \mathcal{M}(X_{t_i,l}) \neq \mathcal{M}(X_{t_j,l})$

Therefore, *completeness* indicates that the whole time period of dataset is covered by the union of all MPB's. *Coherence* expresses the notion that any time points in the same MPB have the identical set of missing features. Finally, *Exclusiveness* suggests that time points in different MPB's must have different sets of missing features.

Definition 6 (Feature Indexing Function)**.** *$\mathcal{W}_m$ is defined as the weight tensor learned by the data for MPB $\{X_{t,l}\}_{t,l}^{T_m,L}$. A feature indexing function $\mathcal{W}_{G(\cdot)}$ is defined as follows:*

$$\mathcal{W}_{G(\cdot)} \equiv \bigcup_m^M [\mathcal{W}_m]_{(\cdot)}$$

For example, $\mathcal{W}_{G(i,j,k)} \equiv \bigcup_m^M [\mathcal{W}_m]_{i,j,k}$ and $\mathcal{W}_{G(i,\cdot,k)} \equiv \bigcup_m^M [\mathcal{W}_m]_{i,\cdot,k}$.

According to Definitions 5 and 6, the feature learning problem based on a 3rd-order strong hierarchy is then formalized as:

$$\begin{aligned}\min_{\mathcal{W}} \mathcal{L}(\mathcal{W}) + \lambda_0 \sum_{i,j,k\neq 0} \|\mathcal{W}_{G(i,j,k)}\|_F + \lambda_1 \sum_{j+k\neq 0} \|\mathcal{W}_{G(\cdot,j,k)}\|_F \\ + \lambda_2 \sum_{i+k\neq 0} \|\mathcal{W}_{G(i,\cdot,k)}\|_F + \lambda_3 \sum_{i+j\neq 0} \|\mathcal{W}_{G(i,j,\cdot)}\|_F \end{aligned} \tag{5.8}$$

where the loss function $\mathcal{L}(\mathcal{W})$ is defined as follows:

$$\mathcal{L}(\mathcal{W}) = \sum\nolimits_{T_m \subseteq T} \frac{1}{|T_m|} \sum\nolimits_{t,l}^{T_m,L} \|Y_{\tau,l} - \mathcal{W}_m \odot Z_{t,l}\|_F^2 \tag{5.9}$$

where $|T_m|$ is the total time period of the MPB T_m.

5.4.3 Model Generalization

The above 3rd-order strong hierarchy-based incomplete feature learning is now extended to Nth-order and prove that the proposed objective function satisfies the Nth-order strong hierarchy. The model is formulated as follows:

$$\min_{\mathcal{W}} \mathcal{L}(\mathcal{W}) + \lambda_0 \sum_{\min(\vec{i})\neq 0} \|\mathcal{W}_{G(\vec{i})}\|_F + \sum_{n=1}^{N} \lambda_n \sum_{\vec{i}_{-n}\neq \vec{0}} \|\mathcal{W}_{G(\vec{i}_{-n})}\|_F \tag{5.10}$$

where $\mathcal{W} = \{\mathcal{W}_m\}_m^M$, and $\mathcal{W}_m \in \mathbb{R}^{|\mathcal{F}_1|\times\cdots\times|\mathcal{F}_N|}$ is an Nth-order tensor whose element index is $\vec{i} = \{i_1, \cdots, i_n\}$. Also denote $\vec{i}_{-n} = \{i_1, \cdots, i_{n-1}, i_{n+1}, \cdots, i_N\}$. $\mathcal{W}_{G(\vec{i})} \equiv \bigcup_m^M [\mathcal{W}_m]_{(\vec{i})}$ according to Definition 6. $\lambda_0 = \lambda/(\prod_i^N |\mathcal{F}_i|)$, $\lambda_n = \lambda/(\sqrt{|\mathcal{F}_n|} \cdot \prod_{i\neq n} |\mathcal{F}_i|)$.

Theorem 3. *The regularization in Equation* (5.10) *enforces a hierarchical structure under an Nth-order strong hierarchy. The objective function in Equation* (5.10) *is convex.*

Proof. First, $\mathcal{L}(\mathcal{W})$ is convex because the Hessian matrix for $\|Y_{\tau,l} - \mathcal{W}_m \odot Z_{t,l}\|_F^2$ is semidefinite. Second, according to Definition 6 and the properties of the norm, $\|\mathcal{W}_{G(\vec{i})}\|_F = \|\bigcup_m^M [\mathcal{W}_m]_{\vec{i}}\|_F$ is convex. Similarly, $\|\mathcal{W}_{G(\vec{i}_{-n})}\|$ is also convex. Therefore, the objective function is convex. □

The proposed model is not restricted to a linear regression and can be extended to generalized linear models, such as logistic regression. The loss function is as follows:

$$\mathcal{L}_M(\mathcal{W}) = -\sum_{T_m \subseteq T} \frac{1}{|T_m|} \sum\nolimits_{t,l}^{T_m,L} \{Y_{\tau,l} \log h(\mathcal{W}_m \odot Z_{t,l}) \\ \cdot (1 - Y_{\tau,l}) \log (1 - h(\mathcal{W}_m \odot Z_{t,l}))\} \tag{5.11}$$

where $h(\cdot)$ could be a nonlinear convex function such as the standard logistic function $h(x) = 1/(1+e^{-x})$.

5.4.4 Parameter Optimization

The problem in Equation (5.10) contains an overlapping group lasso which makes it difficult to solve. To decouple the overlapping terms, an auxiliary variable Φ is introduced and Equation (5.10) is reformulated as follows:

$$\min_{\mathcal{W},\Phi} \ \mathcal{L}_M(\mathcal{W}) + \lambda_0 \sum_{\min(\vec{i})\neq 0} \|\Phi^{(0)}_{G(\vec{i})}\|_F + \sum\nolimits_{n=1}^{N} \lambda_n \sum_{\vec{i}_{-n}\neq \vec{0}} \|\Phi^{(n)}_{G(\vec{i}_{-n})}\|_F \\ s.t. \ \ \mathcal{W}_m = \Phi^{(n)}_m, \ m = 1, \cdots, M; \ n = 1, \cdots, N. \tag{5.12}$$

where the parameter $\Phi^{(n)}_m \in \mathbb{R}^{|\mathcal{F}_1|\times\cdots\times|\mathcal{F}_N|}$ is the auxiliary variable for the mth MPB for Level n. $\Phi_{G(\cdot)}$ then follows Definition 6 such that $\Phi_{G(\cdot)} = \bigcup_m^M [\Phi_m]_{(\cdot)}$. M is defined in Definition 5 and N is the number of levels of the features.

It is easy to see that Equation (5.12) is still convex using Theorem 3. The solution to this constrained convex problem is proposed, which utilizes the alternative direction method of multipliers (ADMM) framework. The augmented Lagrangian function of Equation (5.12) is:

$$L_\rho(\mathcal{W}, \Phi, \Gamma) = \mathcal{L}_M(\mathcal{W}) + \sum\nolimits_{m,n}^{M,N} tr(\Gamma^{(n)}_m (\mathcal{W}_m - \Phi^{(n)}_m)) \\ + \sum\nolimits_{n=1}^{N} \lambda_n \sum_{I_{-n}\neq\vec{0}} \|\Phi^{(n)}_{G(\vec{i}_{-n})}\|_F + \rho/2 \sum\nolimits_{m,n}^{M,N} \|\mathcal{W}_m - \Phi^{(n)}_m\|^2_F \\ + \lambda_0 \sum\nolimits_{\min(\vec{i})\neq 0} \|\Phi^{(0)}_{G(\vec{i})}\|_F \tag{5.13}$$

where ρ is a penalty parameter. $tr(\cdot)$ denotes the trace of a matrix. $\Gamma^{(n)}_m$ is a Lagrangian multiplier for the constraint $\mathcal{W}_m - \Phi^{(n)}_m = 0$.

To solve the objective function in Equation (5.13) with multiple unknown parameters $\mathcal{W}$, Φ, and Γ, hierarchical incomplete feature learning algorithm is proposed as in Algorithm 2. It alternately optimizes each of the unknown parameters until convergence is achieved. Lines 11-12 show the calculation of residuals and Lines 13-19 illustrate the updating of the penalty

parameter, which follows the updating strategy proposed by Boyd et al. [22]. Lines 4-10 show the updating of each of the unknown parameters by solving the subproblems described in the following.

1. Update $\mathcal{W}_m$.

The weight tensor $\mathcal{W}_m$ is learned as follows:

$$\mathcal{W}_m = \underset{\mathcal{W}_m}{\operatorname{argmin}} \mathcal{L}_M(\mathcal{W}) + \frac{N \cdot \rho}{2} \| \frac{1}{N} \sum\nolimits_n^N \Phi_m^{(n)} - \frac{1}{N\rho} \sum\nolimits_n^N \Gamma_m^{(n)} - \mathcal{W}_m \|_F^2 \tag{5.14}$$

which is a generalized linear regression with least squares loss functions. A second-order Taylor expansion is performed to solve this problem, where the Hessian is approximated using a multiple of the identity with an upper bound of $1/(4 \cdot I)$. I denotes the identity matrix.

2. Update $\Phi_m^{(n)}$ $(n \geq 1)$.

The auxiliary variable $\Phi_m^{(n)}$ is learned as follows:

$$\Phi_m^{(n)} \leftarrow \underset{\Phi_m^{(n)}}{\operatorname{argmin}} \frac{\rho}{2} \| \Phi_m^{(n)} - \mathcal{W}_m - \frac{\Gamma_m^{(n)}}{\rho} \|_F^2 + \lambda_n \sum_{\vec{i}_{-n} \neq \vec{0}} | \Phi_{G(\vec{i}_{-n})}^{(n)} \|_F \tag{5.15}$$

which is a regression problem with ridge regularization. This problem can be efficiently using the proximal operator [22].

3. Update $\Phi_m^{(0)}$.

The auxiliary variable $\Phi_m^{(0)}$ is learned as follows:

$$\Phi_m^{(0)} \leftarrow \underset{\Phi_m^{(0)}}{\operatorname{argmin}} \frac{\rho}{2} \| \Phi_m^{(0)} - \mathcal{W}_m - \frac{\Gamma_m^{(0)}}{\rho} \|_F^2 + \lambda_0 \sum_{\min(\vec{i}) \neq 0} | \Phi_{G(\vec{i})}^{(0)} \|_F \tag{5.16}$$

which is also a regression problem with ridge regularization and can be again efficiently solved by utilizing the proximal operator.

4. Update $\Gamma_m^{(n)}$.

The Lagrangian multiplier is updated as follows:

$$\Gamma_m^{(n)} \leftarrow \Gamma_m^{(n)} + \rho(\mathcal{W}_m - \Phi_m^{(n)}) \tag{5.17}$$

5.4.5 Relations to Other Approaches

In this section, it is shown that several classic previous models are actually special cases of the proposed HIML model.

Algorithm 2 Hierarchical Incomplete Feature Learning

Require: Z, Y, λ
Ensure: solution $\mathcal{W}$

Initialize $\rho = 1$, $\mathcal{W}_m, \Gamma, \Phi = \mathbf{0}$.
Choose $\varepsilon_s > 0, \varepsilon_r > 0$.
repeat
  for $m \leftarrow 1, \ldots, M$ **do**
    $\mathcal{W}_m \leftarrow$ Equation (5.14)
    for $n \leftarrow 0, \ldots, N$ **do**
      $\Phi_m^{(n)} \leftarrow$ Equation (5.16) {Equation (5.15) if $n = 0$}
      $\Gamma_m^{(n)} \leftarrow$ Equation (5.17)
    end for
  end for
  $s = \rho \| \{\Phi_m^{(n)} - \Psi_m^{(n)}\}_{m,n}^{M,N} \|_F$ {Calculate dual residual}
  $r = \| \{\mathcal{W}_m^{(n)} - \Psi_m^{(n)}\}_{m,n}^{M,N} \|_F$ {Calculate primal residual}
  if $r > 10s$ **then**
    $\rho \leftarrow 2\rho$ {Update penalty parameter}
  else if $10r < s$ **then**
    $\rho \leftarrow \rho/2$
  else
    $\rho \leftarrow \rho$
  end if
until $r < \varepsilon^r$ and $s < \varepsilon^s$

1. **Generalization of block-wise incomplete multi-source feature learning**. Let $N = 1$, which means there is only one hierarchical level in the multisource data. The proposed model in Equation (5.10) is thus reduced to an incomplete multisource feature learning [108]:

$$\min_W \sum_m \frac{1}{2C_m} \sum_n^{C_m} \|Y_n - W_m \cdot Z_n\|_F^2 + \lambda_0 \sum_i^{|\mathcal{F}|} \|W_{G(i)}\|_F \tag{5.18}$$

where C_m is the count of observations in the mth MPB and $\mathcal{F}$ is the feature set.

2. **Generalization of LASSO**. Let $N = 1$ and $M = 1$, which means there is only one level and there are no missing values. The HIML model is thus reduced to a regression with ℓ_1-norm regularization [79]:

$$\min_W \frac{1}{2C} \sum_i^C \|Y_i - W \cdot Z_i\|_F^2 + \lambda_0 \sum_i^{|\mathcal{F}|} |W_i| \tag{5.19}$$

where C is the count of observations.

3. **Generalization of interactive LASSO**. Let $N = 2$ and $M = 1$, which means there are only 2 hierarchical levels in data without missing value. HIML is thus reduced to a regression

with regularization based on 2nd-order strong hierarchy [46]:

$$\min_{W} \frac{1}{2C} \sum_{i}^{C} \|Y_i - W \odot Z_i\|_F^2 + \lambda_0 \sum\nolimits_{i,j \neq 0} |W_{i,j}| + \lambda_1 \sum\nolimits_{j=1}^{|\mathcal{F}_1|} \|W_{\cdot,j}\|_F + \lambda_2 \sum\nolimits_{i=1}^{|\mathcal{F}_2|} \|W_{i,\cdot}\|_F \tag{5.20}$$

where $\mathcal{F}_1$ and $\mathcal{F}_2$ are the feature sets for the two levels, respectively.

5.5 Experiment

In this section, the performance of the proposed model HIML is evaluated using 10 real datasets from different domains. First, the experimental setup is introduced. The effectiveness and efficiency of HIML is then evaluated against several existing methods for a number of different data missing ratios. All the experiments were conducted on a 64-bit machine with Intel(R) core(TM) quad-core processor (i7CPU@ 3.40GHz) and 16.0GB memory.

5.5.1 Experimental Setup

Datasets and Labels

Table 5.1: Labels of different datasets. (CU=civil unrest; FLU=influenza-like-illnesses).

Dataset	**Domain**	**Label sources** [1]	**#Events**
Argentina	CU	Clarn; La Nacin; Infobae	1306
Brazil	CU	O Globo; O Estado de So Paulo; Jornal do Brasil	3226
Chile	CU	La Tercera; Las ltimas Notcias; El Mercurio	706
Colombia	CU	El Espectador; El Tiempo; El Colombiano	1196
El Salvador	CU	El Diro de Hoy; La Prensa Grfica; El Mundo	657
Mexico	CU	La Jornada; Reforma; Milenio	5465
Paraguay	CU	ABC Color; Ultima Hora; La Nacon	1932
Uruguay	CU	El Pas; El Observador	624
Venezuela	CU	El Universal; El Nacional; Ultimas Notcias	3105
U.S.	FLU	CDC Flu Activity Map	1027

In this chapter, 10 different datasets from different domains were used for the experimental evaluations, as shown in Table 5.1. Among these, 9 datasets were used for event forecasting

[1] In addition to the top 3 domestic news outlets, the following news outlets are included: The New York Times; The Guardian; The Wall Street Journal; The Washington Post; The International Herald Tribune; The Times of London; Infolatam.

Table 5.2: Features of multiple data sources

domain	data sources	features
Civil Unrest	CURRENCY	Open,High,Low,Close
	Tor	Tor network traffic statistics
	ICEWS	CAMEO Codes[8] of event news article content
	Twitter	Volume time series of 982 keywords from [86]
FLU	FluSurv-NET	Influenza Hospitalization Ratio by age groups: 0-4 yr, 5-17 yr, 18-49 yr, 50-64 yr, and 65+ yr
	ILI-Net	weighted/unweighted ILI ratios,positive perctentage, #cases of flu types: A(H1N1), A(N1), A(H3), A, B, H3N2v
	Twitter	Volume time series of 522 keywords from [82]

Table 5.3: Geographical levels and time ranges of the multiple data sources

	Civil Unrest (yyyy-mm-dd)			Influenza (yyyy-week)		
	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
Geo-level	City	State	Country	State	Region	Country
data sources: training period	Twitter: 2013-04-01∼2013-12-31	ICEWS: 2013-04-01∼2013-07-10 2013-10-21∼2013-12-31	CURRENCY: 2013-04-01∼2013-10-21 TOR: 2013-04-01∼2013-10-21	Twitter: 2011-1∼2013-52	ILI-Net: 2009-35∼2013-52	FluSurv-NET: 2009-1∼2011-12 2011-36∼2012-13 2012-36∼2013-52

under the civil unrest domain for 9 different countries in Latin America. For these datasets, 4 data sources from different geographical levels were adopted as the model inputs, which are Twitter, The Onion Router (Tor) network traffic statistics[2], Currency Exchange[3], and Integrated Crisis Early Warning System (ICEWS) counts[4], as shown in Table 5.3. The features of each data source are shown in Table 5.2. The data collected for each source was partitioned into a sequence of date-interval subcollections. The data for the period from April 1, 2013 to December 31, 2013 was used for training, while the data from January 1, 2014 to December 31, 2014, was used for the performance evaluation. The locations of the tweets were all geocoded by the EMBERS geocoder [86]. The event forecasting results were validated against a labeled event set, known as the gold standard report (GSR), exclusively provided by MITRE [74]. GSR is a collection of civil unrest news reports from the most influential newspaper outlets in Latin America [86], as shown in Table 5.1. An example of a labeled GSR event is given by the tuple: (CITY="Hermosillo", STATE = "Sonora", COUNTRY = "Mexico", DATE = "2013-01-20").

The other dataset was collected to track influenza outbreaks in the United States and consists of 3 data sources from different geographical levels, which are Twitter, ILI-Net[5], and FluSurv-

[2]Tor: `https://www.torproject.org/`

[3]Currency Exchange: `http://finance.yahoo.com/currency-converter/`

[4]ICEWS project: `http://www.lockheedmartin.com/us/products/W-ICEWS.html`

[5]`ILI-NET:https://wwwn.cdc.gov/ilinet/`

Table 5.4: Event forecasting performance in civil unrest datasets based on area under the curve (AUC) of ROC

Missing data ratio (3%)									
Method	Argentina	Brazil	Chile	Colombia	El Salvador	Mexico	Paraguay	Uruguay	Venezuela
LASSO	0.5267	0.7476	0.5624	0.8032	0.3148	0.7823	0.5572	0.4693	0.8073
LASSO-INT	0.5268	0.7191	0.5935	0.7861	0.5269	0.777	0.4887	0.5069	0.7543
iMSF	0.4795	0.4611	0.5033	0.7213	0.5	0.5569	0.4486	0.4904	0.5
MTL	0.3885	0.5017	0.5011	0.4334	0.3452	0.4674	0.4313	0.3507	0.5501
Baseline	0.5065	0.7317	**0.6148**	0.8084	**0.777**	**0.8037**	0.7339	0.7264	**0.7846**
HIML	**0.5873**	**0.8353**	0.5705	**0.8169**	0.7191	0.7973	**0.7478**	**0.8537**	0.7488
Missing data ratio (30%)									
Method	Argentina	Brazil	Chile	Colombia	El Salvador	Mexico	Paraguay	Uruguay	Venezuela
LASSO	0.5035	0.7362	0.588	0.8412	0.3785	0.7896	0.478	0.6749	0.681
LASSO-INT	0.4976	0.6361	0.5912	0.8151	0.3852	0.7622	0.426	0.7177	0.6428
iMSF	0.4797	0.4611	0.4959	0.6845	0.5	0.5569	0.4811	0.4898	0.5
MTL	0.4207	0.5156	0.5023	0.5978	0.3413	0.4666	0.4318	0.347	0.4397
Baseline	0.5012	0.7724	**0.6245**	0.8032	**0.7626**	0.7598	0.738	0.8205	**0.7621**
HIML	**0.5854**	**0.8497**	0.6072	**0.8449**	0.726	**0.7907**	**0.7471**	**0.8576**	0.7378
Missing data ratio (50%)									
Method	Argentina	Brazil	Chile	Colombia	El Salvador	Mexico	Paraguay	Uruguay	Venezuela
LASSO	0.5128	0.7461	0.5301	0.8167	0.3139	0.7552	0.5285	0.5992	0.6678
LASSO-INT	0.504	0.6145	0.5537	0.7339	0.4283	0.7309	0.4745	0.5396	0.6155
iMSF	0.4796	0.4611	0.4962	0.7467	0.4899	0.5488	0.4804	0.487	0.5
MTL	0.5104	0.4818	0.4715	0.65	0.3375	0.4744	0.436	0.3578	0.3839
Baseline	0.5101	0.7717	**0.639**	0.8142	**0.7665**	**0.8079**	0.7324	0.8112	**0.7759**
HIML	**0.5795**	**0.8463**	0.548	**0.8432**	0.7126	0.7892	**0.7477**	**0.856**	0.7176
Missing data ratio (70%)									
Method	Argentina	Brazil	Chile	Colombia	El Salvador	Mexico	Paraguay	Uruguay	Venezuela
LASSO	0.5162	0.6674	0.5947	0.8344	0.2597	0.7485	0.4075	0.2652	0.6699
LASSO-INT	0.4691	0.5557	0.5469	0.7167	0.2116	0.7	0.3808	0.2256	0.6503
iMSF	0.4796	0.4611	0.5503	0.7855	0.5	0.557	0.4795	0.5221	0.5
MTL	0.4128	0.5023	0.5069	0.6195	0.3323	0.4702	0.4283	0.3569	0.6464
Baseline	0.5188	0.7741	**0.6059**	0.8121	**0.7557**	**0.8097**	0.7136	0.72	0.6993
HIML	**0.5484**	**0.7812**	0.3887	**0.8416**	0.7181	0.8001	**0.7146**	**0.8453**	**0.716**

NET[6], as shown in Table 5.3. These data sources all have different geographical levels. The features of each data source are shown in Table 5.2. In this case, the data collection for each source was partitioned into a sequence of week-interval subcollections. The data for the period from January 1, 2011 to December 31, 2013 was used for training, while the data from January 1, 2014 to December 31, 2014, was used for the performance evaluation. The locations of the tweets were geocoded by the Carmen geocoder [82]. The forecasting results for the flu outbreaks were validated against the corresponding influenza statistics reported by the Centers for Disease Control and Prevention (CDC)[7]. CDC publishes the weekly influenza-like illness (ILI) activity level for each state in the United States based on the proportional level of outpatient visits to healthcare providers for ILI. There are 4 ILI activity levels: minimal, low, moderate, and high, where the level "high" corresponds to a salient flu outbreak and is effectively the target when forecasting. An example of a CDC flu outbreak event is: (STATE = "Virginia", COUNTRY = "United States", WEEK = "01-06-2013 to 01-12-2013").

[6] FluSurv-NET:http://www.cdc.gov/flu/weekly/overview.htm#Hospitalization

[7] CDC: http://www.cdc.gov/flu/weekly/.

[8] Event data codebook of Conflict and mediation event observations (CAMEO): http://phoenixdata.org/description.

Parameter Settings and Metrics

There is only one tunable parameter in the proposed HIML model, namely the regularization parameter λ. Based on a 10-fold cross validation on the training set, it was set as $\lambda = 0.2$. The logit function was used in Equation (5.11) for the proposed HIML.

In the experiment, the event forecasting task was to predict whether or not there would be an event during the next time step for a specific location. For civil unrest datasets, a time step is one day and the location is a city. For disease outbreaks, a time step is one week and the location is a state. A predicted event was matched to a GSR event if both the time and location attributes were matched; otherwise, it was considered a false forecast. To validate the prediction performance, different metrics were adopted: the True Positive Ratio (TPR) designates the percentage of positive predictions that successfully matched the events that truly happened, while the False Positive Ratio (FPR) denotes the percentage of positive predictions that were actually false alarms. In addition, a Receiver operating characteristic (ROC) curve was utilized to evaluate the forecasting performance as its discrimination threshold for each predictive model was varied. Finally, the use of Area Under ROC Curve (AUC) was also examined as a comprehensive measure of forecasting performance.

5.5.2 Performance

In this section, the effectiveness on the AUC and ROC curves are analyzed for all the comparison methods, including LASSO [79], LASSO with Interactive Features (LASSO-INT), Incomplete Multi-Source Data Fusion (iMSF) [108], Multitask Learning (MTL) [112], and the Baseline.

AUC on civil unrest datasets

Table 5.4 summarizes the effectiveness and robustness comparison for forecasting civil unrest events for different missing data ratios. The AUC measure has been adopted to quantify the performance. The original percentage of missing data in the data sources was 3%. This is manually enlarged to 30%, 50%, and 70% by randomly reducing the number of dates with complete multiple sources.

The results shown in Table 5.4 demonstrate that the methods that take into account the hierarchical topology in the data sources performed better. Specifically, the performance of HIML and the baseline method outperformed the other methods for different missing data ratios. LASSO and LASSO-INT also performed competitively with AUC larger than 0.75 on four datasets. Compared with the other methods, iMSF and MTL had only limited

AccessedFeb2016.

performance for a missing data ratio of 3%. When looking across different missing data ratios, it can be seen that the methods that were best able to handle incomplete input data achieved better robustness against missing values. The performance of LASSO dropped an average 10%, considerably more than iMSF, which dropped less than 3%, when the missing data ratio increased from 3% to 70%. HIML, similar to iMSF, was able to handle the missing value problem in multiple data sources. It also achieved an outstanding model robustness against missing values, dropping on average less than 3% when the missing data ratio increased from 3% to 70%. MTL was also not particularly sensitive to the change in missing values, partially due to its ability to handle the lack of data by sharing the information across different tasks. In all, HIML outperformed all the other methods in 6 out of the 9 datasets for all the different missing data ratios by 6% on average, and achievd the second best performance on the other 3 datasets. This is because HIML effectively handles the two crucial challenges, namely hierarchical topology and interactive missing values.

AUC on the flu dataset

Table 5.5 shows the performance on the metric AUC and training runtime for forecasting influenza outbreaks.

As with the civil unrest datasets, Table 5.5 shows that for the influenza dataset, the methods that take into account the hierarchical topology in the data sources still perform competitively for the missing data ratio of 21% that was present in the real-world dataset. Specifically, the performance of HIML and the baseline method outperformed both iMSF and MTL. LASSO and LASSO-INT also performed competitively, with AUC surpassing 0.85 for different missing data ratios. Compared with the other methods, MTL suffered from a limited performance on a missing data ratio of 21%. When looking across the different missing data ratios, it is apparent that the methods that were best able to handle incomplete input data not surprisingly achieved better robustness against missing values. For example, iMSF performed consistently well, with AUCs between 0.86 and 0.89 even when the missing data ratio increased from 21% to 70% because it was able to cope with the missing value problem in multiple data sources. As with iMSF, HIML also achieved a consistent performance across the full range of missing data ratios. MTL was also not quite as sensitive to changes in the missing data values, which mirrors its performance on the civil unrest datasets, shown in Table 5.4. The performance of the other methods, namely LASSO, LASSO-INT, and Baseline, dropped more significantly. For example, although the Baseline method achieved a good AUC of 0.9044 at a missing data ratio of 21%, this dropped to 0.4359 when the missing data ratio increased to 70% because it could not sufficiently utilize the shared knowledge across different missing patterns and thus large amounts of information were lost. As with the civil unrest datasets, when forecasting influenza outbreaks HIML once again outperformed all the other methods consistently for all the different missing data ratios by clear margins, due to its capacity to handle hierarchical topology and interactive missing data values.

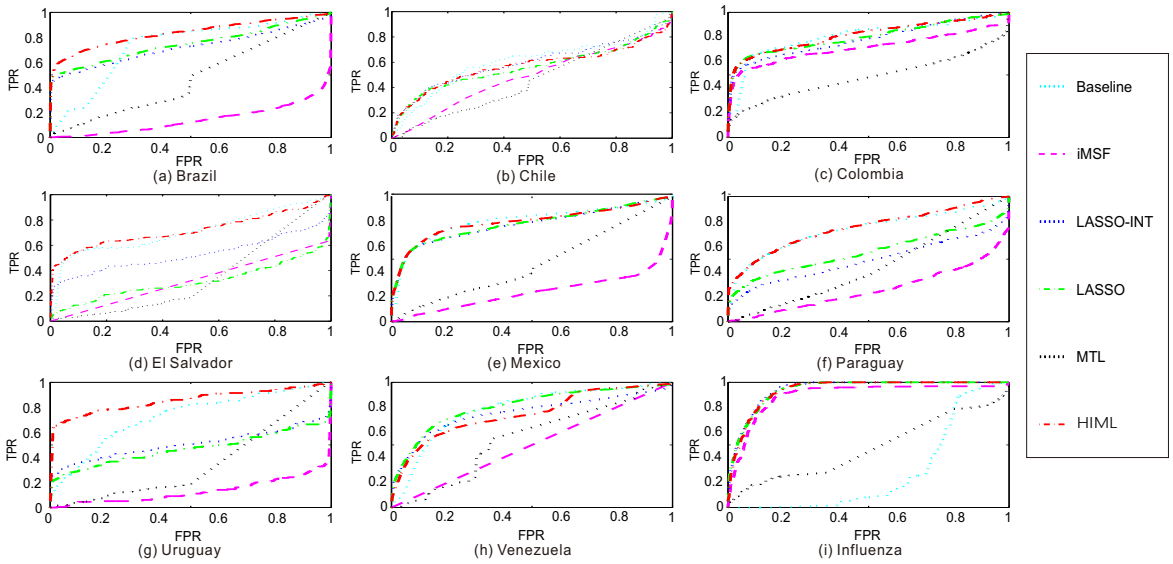


Figure 5.3: Receiver operating characteristic (ROC) curves for the performances on different datasets

Efficiency on running time

The rightmost column of Table 5.5 shows the training time efficiency comparison among HIML and the competing methods for forecasting influenza outbreaks with 21% missing ratio. The running times on the test set for all the comparison methods are instant (i.e., less than 0.01 second for one prediction) so that are not provided here. According to Table 5.5, the running time of the baseline method was 31.97, outperforming the other methods. LASSO, LASSO-INT, MTL, and HIML were hundreds of seconds on the whole training set. However, the running times achieved by all these methods were only a maximum of 15 minutes for a 4-year-long huge training set for week-wise event forecasting tasks, making this eminently practical for real-world applications. The efficiency evaluation results on civil unrest datasets follow a similar pattern of Table 5.5 and are not provided due to space limitations.

Event forecasting performance on ROC curves

Figure 5.3 illustrates the event forecasting performance ROC curves for 9 datasets in two domains, namely civil unrest and influenza outbreaks. The Argentina dataset follows a similar pattern to that of Chile and is thus not shown here to save space. For the 8 civil unrest datasets in Figures 5.3(a)-(h), HIML performs the best overall, with ROC curves covering the largest area above the axis. Moreover, the ROC curves for HIML are consistently above those of the other methods in datasets including Brazil, Colombia, El Salvador, Paraguay, and Uruguay as FPR and TPR vary from 0 to 1. For the datasets for Chile and Mexico, HIML, LASSO, LASSO-INT, and the Baseline perform similarly, all outperforming the other methods. For the dataset for Venezuela, LASSO, LASSO-INT, and the Baseline method

Table 5.5: Event forecasting performance in influenza datasets

	Missing data ratio				runtime
Method	21%	30%	50%	70%	(second)
LASSO	0.9180	0.9056	0.9036	0.8753	493.92
LASSO-INT	0.9142	0.9027	0.9073	0.8403	508.49
iMSF	0.8949	0.8899	0.8930	0.8628	88.90
MTL	0.6129	0.5303	0.6253	0.5568	223.78
Baseline	0.9044	0.9045	0.8562	0.4359	**31.97**
HIML	**0.9372**	**0.9368**	**0.9364**	**0.9357**	851.83

perform best when FPR is smaller than 0.7, while HIML outperforms the other methods when $FPR > 0.7$. MTL generally achieves a limited performance, but its performance is robust against missing ratio, as can be seen in Tables 5.4 and 5.5. For the influenza outbreak dataset, as can be seen from Figure 5.3(i), HIML consistently outperforms the other methods with different FPR and TPR values. iMSF, LASSO, and LASSO-INT also achieve quite competitive performances, outperforming the baseline method and MTL by an apparent margin.

5.6 Conclusions

Significant societal events are prevalent in multiple aspects of society, e.g., economics, politics, and culture. To accommodate all the intricacies involved in the underlying domain, event forecasting should be based on multiple data sources but existing models still suffer from several challenges. this chapter has proposed a novel group-Lasso-based feature learning model that characterizes the feature dependence, feature sparsity, and interactions among missing values. An efficient algorithm for parameter optimization is proposed to ensure global optima. Extensive experiments on 10 real-world datasets with multiple data sources demonstrated that the proposed model outperforms other comparison methods in different ratios of missing values.

Chapter 6

Deep Learning based Epidemics Modeling for Flu Forecasting

This chapter proposes a novel semi-supervised deep learning framework that integrates the strengths of computational epidemiology and social media mining techniques. First, the introduction of research background is introduced in Section 6.1. Section 6.2 reviews existing work in this area. Section 6.3 presents the problem formulation. Section 6.4 elaborates the mathematical descriptions of the SimNest model, and Section 6.5 presents the parameter optimization for SimNest. Section 6.6 introduces the extended functions of SimNest. In Section 6.7, the extensive experimental results are analyzed. This work concludes by summarizing the study's important findings in Section 6.8.

6.1 Introduction

Infectious disease epidemics such as influenza and Ebola pose a serious threat to global public health. According to a recent World Health Organization (WHO) report [102], seasonal influenza alone is estimated to result in about 3 to 5 million cases of severe illness and about 250,000 to 500,000 deaths each year. In the recent Ebola outbreak in West Africa, there have been 27,055 cases and 11,142 deaths [101]. These diseases share two important characteristics: (1) They spread through close contacts between people; With increased local and global travel, the epidemic is often of large spatial scale. (2) They spread rapidly; for example, during the 2009 H1N1 pandemic, the initial case occurred in Mexico in March 2009; but by the beginning of November 2009, more than 6,000 people had died from H1N1 influenza [85]. In order to take effective public health measures to mitigate such fast-developing epidemics, it is crucial to characterize the disease and the evolution of the ongoing epidemic efficiently and accurately. To handle this problem, recent research in both computational epidemiology and social media mining have achieved important progress and demonstrated their respective

usefulness in different aspects.

In the field of computational epidemiology, individual-based network epidemiology has been developed to study the spatio-temporal dynamics of the spread of epidemics. It simulates disease transmission at individual level, and interventions such as vaccinations, school closures, and quarantine. High-performance simulation systems have been developed that are capable of simulating epidemics using network-based models. Such simulations compute the evolution of an epidemic evolution, enabling planners to: (i) forecast the spatio-temporal spread of the disease; (ii) estimate important epidemic measures such as the peak time; and (iii) evaluate the effectiveness of intervention strategies.

Currently, computational epidemiology suffers from the following challenges. 1) *Lack of spatially fine-grained surveillance data for model tuning.* Existing work mostly relies on surveillance data provided by the Centers for Disease Control and Prevention (CDC) [28] to estimate the model parameters. However, CDC surveillance data only provides state-level spatial information, which is insufficient for accurate diffusion modeling within a state. 2) *Difficulties in tracking the dynamics of contact networks in real time.* Intervention, such as school closures and vaccinations play an important role in mitigating epidemics by changing people's infectivity and vulnerability and altering the contact network structure. Current approaches lack effective mechanisms to monitor the impact of ongoing interventions during the current season in real time. 3) *High cost and low timeliness of retraining.* Existing approaches generally rely on batch training based on the CDC surveillance data. However, CDC surveillance data is updated weekly, with a delay of at least one week, and thus cannot catch up with the real time disease spread.

Social media, on the other hand, can capture timely and ubiquitous disease information from social sensors (i.e., social media users) [33]. Social media-based approaches can be classified into two categories: (i) aggregate-level disease surveillance and (ii) detailed health-informatics analysis. The first category assumes that self-reported symptoms from social media users are reliable signals reflecting the aggregate-level trend of a particular outbreak. Among these, some focus on detecting or tracking current influenza outbreaks while others aim to forecast the severity of the outbreak. The second category focuses on detailed modeling of the social media contents as well as their relevance to health informatics, disease geoinformatics, and health behaviors. However, social media mining approaches suffer from three major drawbacks. First, as a crucial determinant of the disease diffusion pattern, real contact networks are basically unobservable. Estimating social contact networks merely based on the location of social media users is neither accurate nor sufficient. Second, they generally can only characterize the health information of social media users, but not the whole demographic population. Third, they typically only employ the disease information retrieved from social media without utilizing disease model knowledge.

Although computational epidemiology can model the progress of a disease and the underlying disease contact network among individuals, it suffers from a lack of timely and fine-grained surveillance data. Social media mining, on the other hand, provides spatiotemporal surveil-

lance with good timeliness and geographical details, but is unable to observe the underlying contact network and disease progress model. In order to overcome the above-mentioned challenges, this chapter proposes a novel online semi-supervised deep learning framework that integrates the strengths of individual-based epidemic simulation and social media mining techniques, named **S**oc**I**al **M**edia **N**ested **E**pidemic **S**imula**T**ion (**SimNest**). SimNest is a novel bispace framework that combines computational epidemiology and social media data by an interactive mapping, as shown in Figure 6.1. Specifically, on one hand, the health states and interventions actions of social media users are not only identified via their posts by deep learning, but also are regularized unsupervisedly by the disease model in computational epidemiology. On the other hand, the user health states and parametrized disease model learned from social media can provide the computational epidemic model with individual-level surveillance and the optimized disease model parameters. This interactive learning process between social media and computational epidemiolgoy iteratively performs, leading to a consistent stage between these two spaces. The main contributions of this study are summarized as:

- **Proposing a novel integrated framework for computational epidemiology and social media mining**: The existing approaches from computational epidemiology and social media mining focus on different but complementary aspects. The former focuses on modeling the underlying mechanisms of disease diffusion while the latter provides timely and detailed disease surveillance. SimNest framework utilizes both type of information by integrating the strengths of them.
- **Developing a semi-supervised multilayer perceptron (MLP) for mining epidemic features**: To achieve deep integration, this work enforces unsupervised pattern constraints derived from epidemic disease progress model onto the supervised classification. Using this semi-supervised strategy, the sparsity of labeled data can be solved.
- **Designing an online training algorithm**: To minimize the inconsistencies between Twitter space and the simulated world, this chapter proposes to iteratively optimize model parameters via an online algorithm. This algorithm ingests the social media data streams and updates the model parameters in real time, which not only reduces the cost of retraining but also ensures the timeliness of the model.
- **Conducting extensive experiments for performance evaluations**: The proposed SimNest model was evaluated using Twitter data collected from Jan 2011 to Apr 2015 in 4 states and the District of Columbia in the United States. The proposed methods consistently outperformed competing methods in multiple metrics. The advantage of integrating the complementary strengths of computational epidemiology and social media mining is demonstrated.

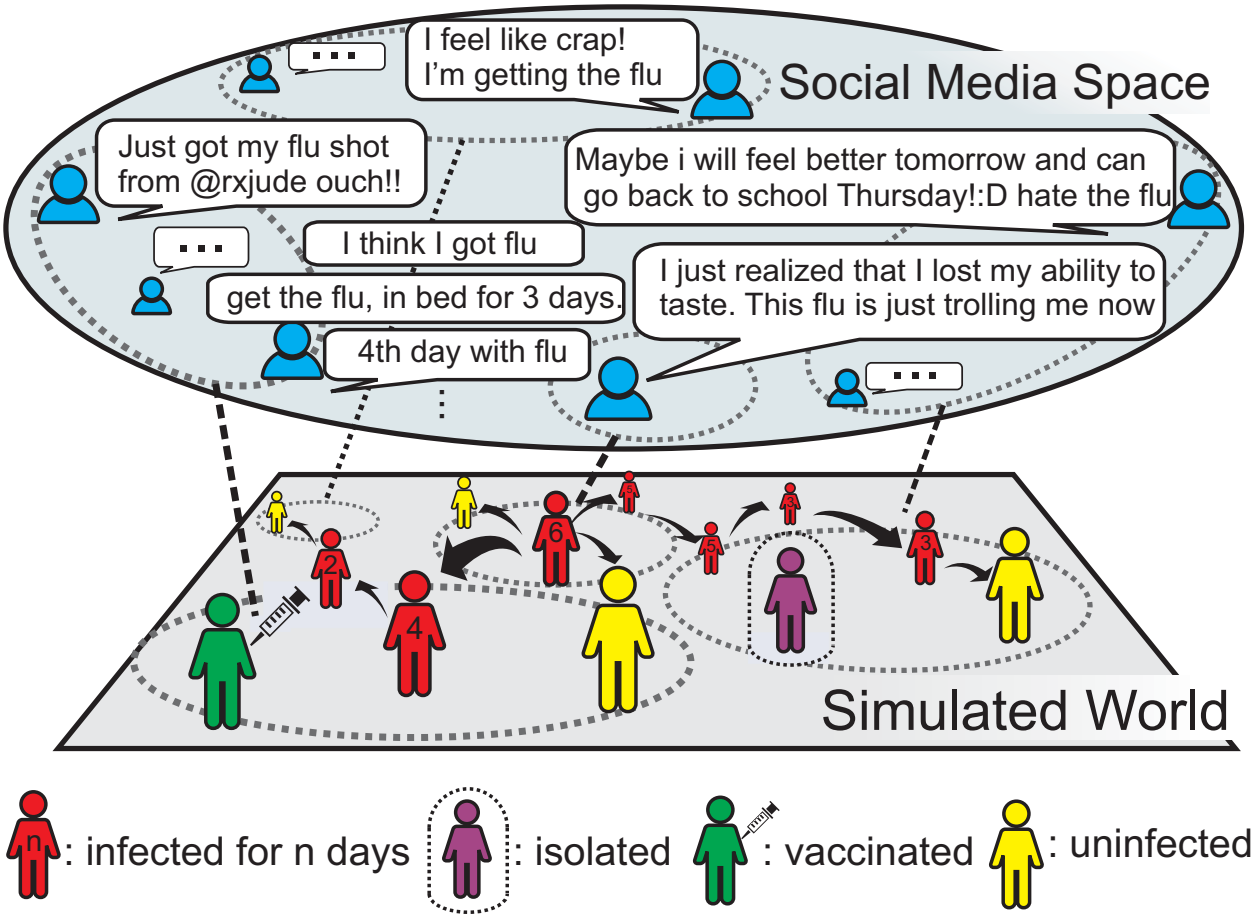


Figure 6.1: In SimNest, the simulated world mirrors social media space. The posts of social media users reflect their statuses information of health, vaccination, or isolation. This information is mapped to the corresponding spatial subregions in the demographics-based contact network in the simulated world.

6.2 Related Work

Computational models for epidemiology are important for a number of reasons. Traditionally computational epidemiology focused on *compartmental models*, where a population is divided into subgroups (compartments) based on people's health state and demographics, and the epidemic dynamics are modeled by ordinary differential equations [76, 97].

Recently, individual-based computational models have been developed to support network epidemiology, where an epidemic is modeled as a stochastic propagation over an explicit interaction network between people. One common approach taken by network epidemiology is to model the interactions between people using random graph models [36, 44]. Here, the closed form analytical results obtained can be applied to study epidemic dynamics, but this relies on the inherent symmetries in random graphs. With no explicit location modeling, it cannot be applied to compute the geographical spread of an epidemic.

Another direction taken by network epidemiology is to develop a realistic representation of a population by considering members' social contact network, and then using individual-based simulations to study the spread of epidemics in the network [12,17]. This approach first constructs a synthetic population, where each individual is assigned demographic, geographic, social, and behavioral attributes so that at various aggregate levels the synthetic population is statistically indistiguishable from the real population. The synthetic individuals are also assigned daily activities and their physical locations at any moment, so by connecting all persons located within close proximity to each other one can construct the corresponding synthetic social contact network for the population [11]. Individual-based simulations model epidemics as diffusion processes across this network, and compute who infects whom at what time at which location [17]. In addition to the synthetic network and disease model, another key component of individual-based epidemic simulations is the associated public health and individual interventions, which can be either pharmaceutical such as vaccination, or non-pharmaceutical such as social distancing. These interventions affect the epidemic evolution by changing the node or edge properties of the network.

Recently, there have been a number of proposals for influenza epidemic knowledge mining techniques based on social media, which can be categorized into two threads. The first thread focuses on *aggregate level disease surveillance*. For example, Krieck et al. [58] suggested that self-reported symptoms are the most reliable signal in detecting whether a tweet is relevant to an outbreak or not and then went on to demonstrate that this is because even though people generally do not identify their specific problem until diagnosed by an expert, they readily write about how they feel. Using a similar approach to identify flu-related tweets, researchers generally concentrated on tracking the overall trend of a particular disease outbreak, typically influenza, by monitoring social media [3, 37, 51, 110].

The second thread focuses on *detailed health-informatics semantic analysis*. These approaches typically model the language of the social media messages and their relevance to public health [81] influenza surveillance [35], disease geoinformatics [38], user interac-

tions [23], and health behavior [33]. Paul et al. [81] proposed a topic model that captures the symptoms and possible treatments for ailments, and then went on to propose a way to identify the geographical patterns in the prevalence of such ailments. Specific to self-reporting on influenza, Collier et al. [35] categorized five sub-classes of tweets that serve as user behaviour response surveys for influenza outbreaks, Dredze et al. [38] focused on achieving accurate geographical location identification for influenza outbreak detection, Brennan et al. [23] utilized Twitter user interactions to uncover the health condition of Twitter users. Tackling the problem from a different direction, Chen et al. [33] concentrated on modelling the disease progression in individuals.

6.3 Problem Setup

This work aims to characterize the spatiotemporal diffusion of epidemics across the underlying social contact network. Specifically, assume the discrete time increases by interval, and there are T such time intervals $\mathcal{T} = \{0, \cdots, t, \cdots, T\}$. This chapter aims to know for each time interval $t \in \mathcal{T}$ the health states $\mathcal{Z}$ of the people in the population. Regarding health state transition in a time interval t, this chapter does not distinguish between different moments during the interval when it occurs exactly. To address this problem, approaches based on computational epidemiology and social media mining are formulated in turn below.

6.3.1 Individual-based Epidemic Simulation

A disease transmits through people to people contacts. These people-people contacts form a network called a social contact network $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, which is a directed, edge-weighted network. Nodes $\mathcal{V}$ correspond to individuals in the population. An edge $(v_1, v_2) \in \mathcal{E}$ with weight $\mathcal{W}(v_1, v_2)$ denotes the nodes v_1 and $v_2 \in \mathcal{V}$ has a contact of duration $\mathcal{W}(v_1, v_2)$. During the contact the disease may transmit from node v_1 to v_2 with probability $p(\mathcal{W}(v_1, v_2), \tau)$, where τ, called transmissibility, is probability of transmission per unit of contact time and is a parameter associated with the disease. It is first assumed that the contact network $\mathcal{G}$ is constant. In Section 6.6, the situation when $\mathcal{G}$ changes with interventions will be considered.

Each person is assumed to be in one of the following four health states at any time: *susceptible (S), exposed (E), infectious (I), and recovered (R)*, which is known as the SEIR disease model. It is widely used in the mathematical epidemiology literature [6, 76]. Associated with each person v are an incubation period $p_E(v)$ and an infectious period $p_I(v)$, each from a distribution. It is assumed that both are normally distributed, i.e., $p_E(v) \sim \mathcal{N}(\mu_E, \sigma_E)$ and $p_I(v) \sim \mathcal{N}(\mu_I, \sigma_I)$. A person is in the susceptible state until he becomes exposed. If a person v becomes exposed, he remains so for $p_E(v)$ days, during which he is not infectious. Then he becomes infectious and remains so for $p_I(v)$ days. Finally he recovers and remains so. The transition $S \mapsto E$ is probabilistic. But it is assumed that

once person v becomes exposed, $p_E(v)$ and $p_I(v)$ are sampled from the two normal distributions respectively so their values are determined. In sum, given the parameters, let $Z_{v,t}(p_E(v), p_I(v)) \in \{S, E, I, R\}$ denote the health state of person $v \in \mathcal{V}$ on time $t \in \mathcal{T}$. Therefore, we have $\mathcal{Z} = \{Z_{v,t}(p_E(v), p_I(v))\}_{v \in \mathcal{V}, t \in \mathcal{T}}$, where $\mathcal{Z}$ stands for peoples' inferred health states based on individual-based epidemic simulations.

6.3.2 Social Media Based User Health State Inference

Social media is a popular way for people to post about their everyday feelings, and is commonly treated as a surrogate for the physical world [3]. Taking Twitter as an instance, suppose the set of Twitter users who have ever mentioned their flu infectiousness is denoted as $\mathcal{U} \subseteq \mathcal{V}$, which can increase with Twitter data streams. Each user posts $n_{u,t}$ tweets in each time interval t (e.g., hour, day), $t = 1, 2, \cdots, T$. Define Twitter streams as $\mathcal{D} = \{D_{u,t}\}_{u \in \mathcal{U}, t \in \mathcal{T}}$, where the matrix $D_{u,t} \in \mathbb{Z}^{|V| \times n_{u,t}}$ denotes the posts from user u in time t. The (i, j)-th entry, denoted as $D_{u,t,i,j}$, refers to the frequency of the i-th term in the j-th tweet. V refers to the vocabulary. Suppose we have a predefined subset of keywords $\mathcal{K}$ related to flu, and denote A as the corresponding incidence matrix, $A \in [0, 1]^{|\mathcal{K}| \times |V|}$. Define a matrix $X_{u,t}$ as follows: $X_{u,t} = A \cdot D_{u,t} \cdot \mathbf{1}$, where $\mathbf{1}$ denotes a vector of all ones. It is clear that $X_{u,t} \in \mathbb{Z}^{|\mathcal{K}| \times 1}$ is the vector of keywords frequencies from user u at time t. Hence, $X_u = \{X_{u,t}\}_t^{\mathcal{T}}$ denotes the keyword vectors of user u, while $\mathcal{X} = \{X_u\}_{u \in \mathcal{U}}$ denotes the set of all the keyword vectors. We are interested in learning a classifier f_W, which maps the social media user textual content $X_{u,t}$ to their corresponding health states $Y_{u,t}$:

$$f_W(X_{u,t}) : X_{u,t} \to Y_{u,t} \tag{6.1}$$

where $Y_{u,t} = \mathbf{1}[Z_{u,t} = I]$, I stands for "Infectious", and $\mathbf{1}[\cdot]$ stands for the indicator function. Therefore, $Y_{u,t} = 1$ signifies that user u's health state $Z_{u,t}$ at time t is infectious (I); and $Y_{u,t} = 0$ that it is not. $Y_u = \{Y_{u,t}\}_t^{\mathcal{T}}$ denotes all the health states of user u. W denotes the parameter set of the classifier.

There are three main challenges when using either individual-based epidemic simulation or social media mining techniques individually: (1) There is as yet no surveillance data that is sufficiently real-time and fine-grained to permit the detailed progress of the epidemic simulation to be linked consistently with the physical world. (2) The people-people disease contact network and disease model is hidden to social media data. (3) The fast-streaming and time-evolving nature of huge social media data requires efficient updating of the trained model. Traditional batch-based training suffer from high expense and poor timeliness.

In order to overcome the above-mentioned challenges in either of the above threads individually, this work proposes using both types of information by deeply integrating the strengths of individual-based epidemic simulation and social media mining techniques in the proposed new framework, **S**oc**I**al **M**edia **N**ested **E**pidemic **S**imula**T**ion (**SimNest**), which is elaborated in the following section.

6.4 SimNest Model

As shown in Figure 6.2(A), SimNest learns the users' health states from social media posts based on a multilayer feature representation. Other than considering each time point individually, SimNest utilizes disease progress model in computational epidemiology to constrain the temporal pattern of health states in two aspects: (1) constraining the infectious period to follow a probability distribution in Figure 6.2(C) and (2) resisting a temporally discontinuous health states like in Figure 6.2(D). As shown in Figure 6.2(B), by mapping social media users' health states into demographics-based synthetic contact network, an interactive learning between these two spaces is achieved. Specifically, simulation model parameters are adjusted by the social media surveillance data while the weights of the multilayer-based health state model are regularized by the underlying synthetic disease contact network.

To make the underlying health states in the contact network $\mathcal{G}$ consistent with those gathered from social media data D, SimNest simultaneously optimizes contact network, disease progress model parameters p_I and p_E, and social media-based health state inference $f_W(\cdot)$. Among all the keyword vectors $\mathcal{X}$, we are given a set of labeled samples $\mathcal{X}_1 = \{X_{u,t}\}_{u\in\mathcal{U}_1,t\in\mathcal{T}}$ with corresponding class label $\mathcal{Y}_1 = \{Y_{u,t}\}_{u\in\mathcal{U}_1,\mathcal{T}}$, and unlabeled samples $\mathcal{X}_2 = \{X_{u,t}\}_{u\in\mathcal{U}_2,t\in\mathcal{T}}$, where $\mathcal{U}_2 = \mathcal{U} - \mathcal{U}_1$ is the set of all the unlabeled users. Mathematically, SimNest model is formulated as jointly minimizing the following four loss functions: (A) Supervised loss, (B) Bispace consistency loss, (C) Infectious duration loss, and (D) Temporal proximity loss, as illustrated as below.

$$\begin{aligned}\mathcal{L} = \mathcal{L}_1(\mathcal{Y}_1, \mathcal{X}_1, W) + \mathcal{L}_2(\mathcal{X}_2, \mathcal{G}, p_E, p_I, W) \\ + \mathcal{L}_3(\mathcal{X}_2, p_I, W) + \mathcal{L}_4(\mathcal{X}_2, W)\end{aligned} \tag{6.2}$$

The different loss functions are illustrated in Figure 6.2. In the following subsections, each of these will be elaborated.

6.4.1 Supervised Loss

The mapping $f_W(\cdot)$ between tweet texts and user health states is substantialized by applying deep data representation, namely multilayer perception:

$$\begin{aligned}f_W(x) = s(h^{(1)}) = s(\sum\nolimits_{j=1}^{m} W_j^{(2)} s(h_j^{(2)}) + W_0^{(2)}), \\ h_j^{(2)} = \sum\nolimits_{i=1}^{|\mathcal{K}|} W_{j,i}^{(1)} x_i + W_{j,0}^{(1)}\end{aligned} \tag{6.3}$$

apart from the input layer that is the tweet text and the output layer that is the user health state, another hidden layer represents the abstract semantics, where m is the number of hidden layer features. $W = W^{(1)} \cup W^{(2)}$, where $W^{(1)} \in \mathbb{R}^{|\mathcal{K}|\times m}$ is the weight matrix for the mapping from text layer to abstract semantics layer, $W^{(2)} \in \mathbb{R}^{m\times 1}$ is the weight vector for

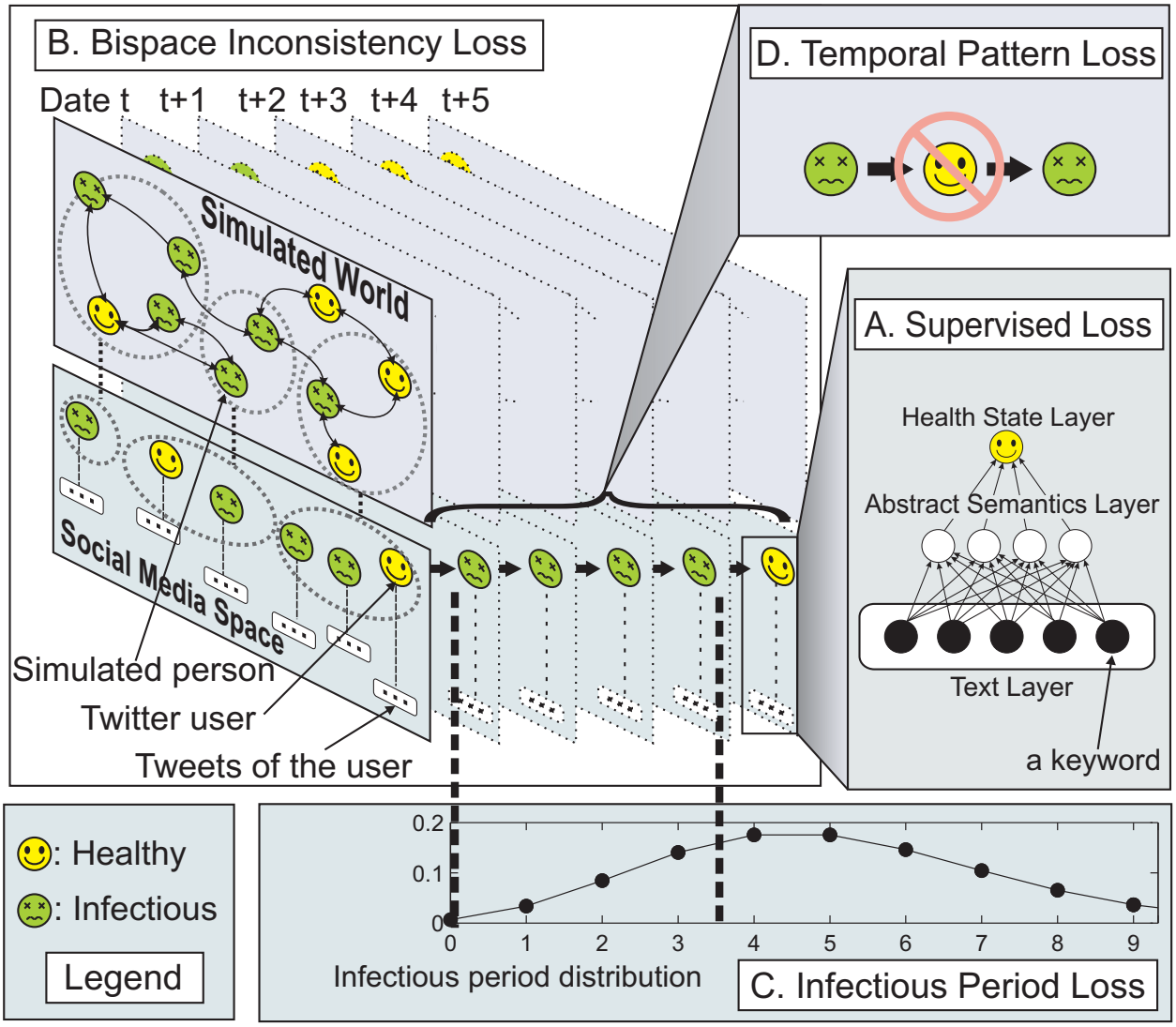


Figure 6.2: The illustration of the SimNest model.

the mapping from abstract semantics layer to the user health status layer and $s(\cdot)$ is the sigmoid function. $h^{(1)} = \sum_{j=1}^{m} W_j^{(2)} s(h_j^{(2)}) + W_0^{(2)}$.

A common way to learn W is to define a loss function over the training data, and then obtain the best W by minimizing the loss of misclassification towards labels:

$$\mathcal{L}_1 = \min_{W} \sum_{u}^{\mathcal{U}_1} \sum_{t}^{\mathcal{T}} \|f_W(X_{u,t}) - Y_{u,t}\|^2 \tag{6.4}$$

6.4.2 Bispace Consistency Loss

To sufficiently benefit from the complementary advantages of individual-based epidemic simulation and social media data, the inner inconsistency of the integrated model need to be minimized. Specifically, the hidden health states in the individual-based epidemic simulation need to be consistent with the observations from social media. On the other hand, the intelligence gleaned from the social media data also needs to correspond to the hidden disease progression across the hidden contact network. More formally, the goal here is formulated as the following loss function:

$$\mathcal{L}_2 = \min_{\Theta,W} \sum_{v}^{\mathcal{V}} \sum_{t}^{\mathcal{T}} \|Q_{v,t}(\mathcal{G}, p_E, p_I) - f_W(X_{v,t})\|^2 \tag{6.5}$$

where $Q_{v,t}(\mathcal{G}, p_E, p_I) = \mathbf{1}[Z_{v,t}(p_E(v), p_I(v)) = I]$, and I stands for the state of "infectious", as introduced in Section 6.3. $\Theta = \{\mathcal{G}, p_E, p_I\}$ are the parameters of individual-based epidemic simulaiton and $p_E(v) \sim \mathcal{N}(\mu_E, \sigma_E)$ and $p_I(v) \sim \mathcal{N}(\mu_I, \sigma_I)$ are the incubation and infectious duration distributions of person v, respectively.

But it is impossible to link the corresponding person to a specific user in Twitter, and not all the people post tweets. Fortunately, however, the specific spatial subregion (e.g., blocks, counties, etc.) of Twitter user $u \in \mathcal{U}$ and simulated individual $v \in \mathcal{V}$ can be known. Hence, the above loss function can be resorted to a fine-grained spatial subregion:

$$\mathcal{L}_2 = \min_{\Theta,W,\lambda_1} \sum_{l,t}^{L,\mathcal{T}} \left\| \lambda_1 \sum_{v}^{\mathcal{V}_l} Q_{v,t}(\mathcal{G}, p_E, p_I) - \sum_{u}^{\mathcal{U}_{2,l}} f_W(X_u, t) \right\|^2 \tag{6.6}$$

where $\mathcal{U}_{2,l}$ denotes the Twitter users in location l, $\mathcal{V}_l$ denotes the people in location l, and λ_1 is the parameter scaling the person count in the individual-based epidemic simulation down to the count of social media users .

6.4.3 Infectious Period Loss

Existing social media mining techniques typically do not assume a specific disease progression model and hence cannot take advantage of its important knowledge pattern. Unlike them,

SimNest borrows the disease progression model from the epidemic simulation to regularize the patterns in the huge unlabeled social media data. This not only greatly mitigates the problem of label data sparsity, but also improves the timeliness and generalization of the modeling. Specifically, the infectious duration of a Twitter user is dependent on the flu outbreak's characteristics as well as his or her general state of physical health, denoted as the following normal distribution:

$$[\sum_{t}^{\mathcal{T}} f_W(X_{u,t})] = d_u \sim p_I(u) = \mathcal{N}(u|\mu_I, \sigma_I) \tag{6.7}$$

By maximizing the likelihood function for the observations, the following objective function is obtained:

$$\max \prod_{u}^{\mathcal{U}_2} N(d_u|\mu_I, \sigma_I) = \max \sum_{u}^{\mathcal{U}_2} \log N(\sum_{t}^{\mathcal{T}} f_W(X_{u,t})|\mu_I, \sigma_I)$$

which can be transformed to the following formulation by considering Equation 6.1:

$$\mathcal{L}_3 = \min_{W,p_I} \frac{1}{2\sigma_I^2} \sum_{u}^{\mathcal{U}_2} \left\| \sum_{t}^{\mathcal{T}} f_W(X_{u,t}) - \mu_I \right\|^2 + \frac{|\mathcal{U}_2|}{(2\pi\sigma_I^2)^{1/2}} \tag{6.8}$$

6.4.4 Temporal Proximity Loss

Another important intrinsic pattern in the health state modeling is that the states in the neighboring time points should be similar. Moreover, a person recovering from the flu typically cannot get the flu again in the same flu season, as illustrated in Figure 6.2(D). Thus, the infectious dates are temporally consecutive. This fact motivates the loss function for the proximity of the neighbor states:

$$\mathcal{L}_4 = \min_{W} \sum_{u}^{\mathcal{U}_2} \sum_{t}^{\mathcal{T}} \|f_W(X_{u,t}) - f_W(X_{u,t+1})\|^2 \tag{6.9}$$

6.5 Online Training Algorithm

To efficiently solve the optimization problem presented in Equation 6.2, this work proposes an online parameter optimization framework. It adopts an alternating minimization approach, where all variables are fixed except for the one being updated.

6.5.1 Solving for W

The process of solving W is based on stochastic gradient descent (SGD) [16]. Training with SGD makes it possible to handle very large databases since every update involves one (or a

pair) of examples, and grows linearly in time with the size of the dataset. The convergence of the algorithm is also ensured for low enough values of threshold error.

The derivatives of $\mathcal{L}_1, \mathcal{L}_3, \mathcal{L}_3$, and $\mathcal{L}_4$ can be deduced using backpropagation algorithms and its variants.

6.5.2 Solving for Θ

Solving for $\Theta = \{\mathcal{G}, p_E, p_I\}$ with respect to the loss function $\mathcal{L}_2$ is a nonconvex and non-differentiable problem, so a numerical optimization algorithm such as the Nelder-Mead method [16] can be adopted to solve it.

6.5.3 Solving for p_I, λ_1

The sufficient statistics μ_I and σ_I of the infectious period distribution p_I have the following analytical solution:

$$\mu_I = \frac{1}{|\mathcal{U}_2|} \sum_{u}^{\mathcal{U}_2} \sum_{t}^{\mathcal{T}} f_W(X_{u,t}) \tag{6.10}$$

$$\sigma_I = (\sum\nolimits_{u}^{\mathcal{U}_2} \sum\nolimits_{t}^{\mathcal{T}} f_W(X_{u,t} - \mu_I)/|\mathcal{U}_2|)^{1/2} \tag{6.11}$$

Solving for λ_1 according to the loss function $\mathcal{L}_2$ in Equation 6.6 yields the following analytical solution:

$$\lambda_1 = \sum_{l,t}^{L,\mathcal{T}} \sum_{u}^{\mathcal{U}_{2,l}} f_W(X_u, t) / \sum_{l,t}^{L,\mathcal{T}} \sum_{v}^{\mathcal{V}_l} Q_{v,t}(\mathcal{G}, p_E, p_I) \tag{6.12}$$

Utilizing the above alternating optimization process, SimNest is trained and utilized to forecast the spatiotemporal epidemic diffusion progress in the online fashion illustrated in Algorithm 3. Specifically, the unlabeled data set $\mathcal{X}$ is continually updated by the social media data streams, with the most out-dated data (such as three months old) being replaced by the newly-arriving data. Then, the weight matrix W is optimized via a SGD fashion until convergence. Utilizing the optimized infectious period distribution as the input for the simulation process, the epidemic simulation parameter p_E is optimized by minimizing the inconsistencies with social media data. Finally, the population's health status $\mathcal{Z}$ is predicted. The optimized parameter p_E is then utilized for the next-step's optimization of weight matrix W with the updated unlabeled data. Therefore, as the data is streaming, the parameters is being optimized with the newest data and the predicted health states $\mathcal{Z}$ streams out.

ALGORITHM 3: Online Algorithm for SimNest

Input: Data matrix $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$, Twitter data stream $\mathcal{C}$, contact network $\mathcal{G}$.
Output: the population's predicted health states $\mathcal{Z}$.

Set the learning rate $\eta = 0.5$. Initialize weight matrix W as matrix of random values between -1 and 1;
repeat
  Update unlabeled data set $\mathcal{X}_2$ by Twitter data stream;
  repeat
    Randomly select a labeled sample $(X_{u,t}, Y_{u,t})$;
    $W \leftarrow W - \eta \cdot \frac{\partial \mathcal{L}_1(X_{u,t}, Y_{u,t}, W)}{\partial W}$;
    Randomly select an unlabeled sample X_u;
    $W \leftarrow W - \eta \cdot \frac{\partial \mathcal{L}_3(X_u, p_I, W)}{\partial W}$;
    Randomly select an unlabeled sample X_v;
    for $i \leftarrow 1$ **to** T **do**
      $W \leftarrow W - \eta \cdot \frac{\partial \mathcal{L}_4(X_{v,i}, X_{v,i+1}, W)}{\partial W}$
    end
    Randomly select a user u from a location $l \in L$;
    $W \leftarrow W - \eta \cdot \frac{\partial \mathcal{L}_2(X_{u,t}, \mathcal{G}, p_E, p_I, W)}{\partial W}$;
    $\mu_I \leftarrow \frac{1}{|\mathcal{U}_2|} \sum_{u}^{\mathcal{U}_2} \sum_{t}^{\mathcal{T}} f_W(X_{u,t})$;
    $\sigma_I \leftarrow (\sum_{u}^{\mathcal{U}_2} \sum_{t}^{\mathcal{T}} f_W(X_{u,t} - \mu_I)/|\mathcal{U}_2|)^{1/2}$;
  until *converge*;
  $p_E, \mathcal{Z} \leftarrow \min \sum_{t}^{\mathcal{T}} \sum_{l}^{L} \left\| \lambda_1 \sum_{v}^{\mathcal{V}_l} Q_{v,t}(\mathcal{G}, p_E, p_I) - \sum_{u}^{\mathcal{U}_{2,l}} f_W(X_{u,t}) \right\|^2$;
  $\lambda_1 \leftarrow \sum_{l,t}^{L,\mathcal{T}} \sum_{u}^{\mathcal{U}_{2,l}} f_W(X_u, t) / \sum_{l,t}^{L,\mathcal{T}} \sum_{v}^{\mathcal{V}_s} Q_{v,t}(\mathcal{G}, p_E, p_I)$
until *the end of data stream*;

6.6 Extensions

6.6.1 Dynamics of Contact Network

In the epidemic diffusion progression, interventions are among the most common and effective ways for the government and individuals to reduce the potential impact from disease outbreaks. Interventions influence the epidemic diffusion largely by changing the people-people contact network. They can be categorized into two types: (1) Pharmaceutical (PI) versus (2) Non-pharmaceutical (NPI). PI interventions, such as administering antivirals and vaccines, can change the characteristics (e.g., disease transmissibility) of the person nodes in the social contact network, while NPI interventions are those actions that effectively change the contact network structure, including school closures, quarantine and sequestration. Therefore,

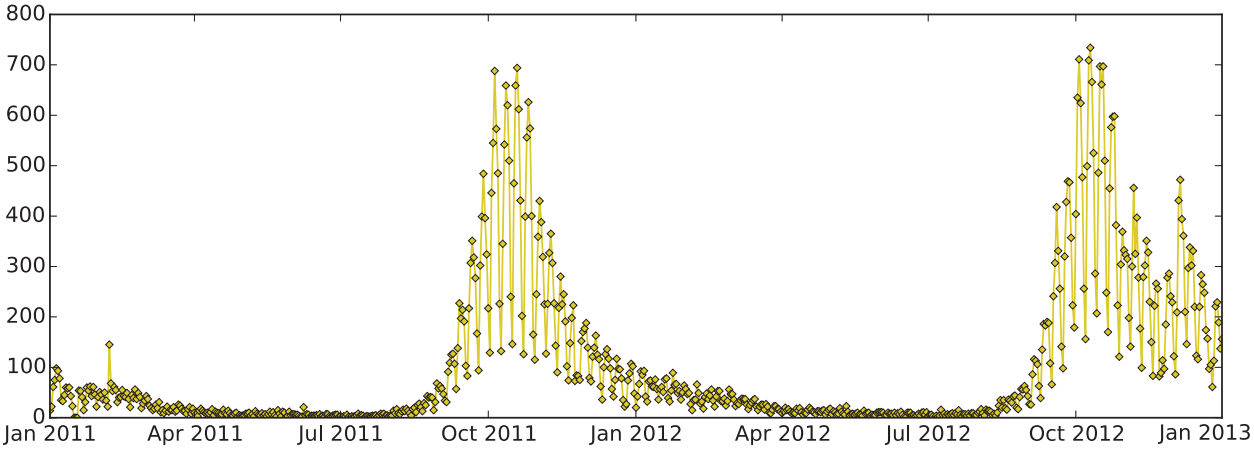


Figure 6.3: Counts of Twitter users in Virginia who got flu shot

both types of interventions can result in changes in the social contact network.

The SimNest framework accommodates these heterogeneous dynamics of contact network effectively via two aspects: (1) Timely intervention actions monitoring based on social media data; and (2) Intervention substantialization through the epidemic simulation process. Take vaccination as an example. First, tweets like "I just got flu shot, it still hurts." that mention their user $\mathcal{U}_l$'s vaccinations from each subregion $l \in L$ are identified by the text classifiers. In the experiments, a 78% identification accuracy is achieved based on the cross-validation results. For example, Figure 6.3 shows the users who got the flu shots as identified by their Twitter postings during Jan 2011 and Jan 2013 in Virginia. It clearly demonstrates both yearly and weekly periodicity and a peak time around November of each year. The relative vaccination ratio in different subregions can then be estimated as $r_l = |\mathcal{U}_l|/\lambda_1|\mathcal{V}_l|$, where $|\mathcal{V}_l|$ is the size of the population in subregion l and λ_1 is the population size scaling factor from the physical world to the Twittersphere, as calculated by Equation 6.12. Next, in the epidemic simulation SimNest substantializes the vaccinations by reducing the transmissibility $p(\mathcal{W}(v_1, v_2))$, ($v_1 \in \mathcal{V}_l$ or $v_2 \in \mathcal{V}_l$) of $r_l \cdot |\mathcal{V}_l|$ random individuals in region l by a ratio, which can either be set by domain knowledge or literature.

6.6.2 Heterogeneous Surveillance Data

The SimNest framework is also flexible to involve multiple surveillance data sources. In the basic problem definition, social media data is merely utilized as a fine-grained surveillance data. Other than that, SimNest allows the addition of heterogeneous surveillance data sources such as CDC [28] surveillance data for the United States, and Paho [80] surveillance data for Latin America. Take CDC surveillance data as an instance, because it is state-level weekly aggregate data, to be comparable to it, SimNest aggregates the predicted user health states into state-level weekly data and involves the following loss function into Equation 6.2,

and get the following equation:

$$\mathcal{L}_c = \min_{W,\lambda_2} \sum_{i}^{T'} \|\lambda_2(a_e - a_s + 1) \sum_{l,t=a_s}^{L,a_e} \sum_{u}^{\mathcal{U}_{2,l,t}} f_W(X_{u,t}) - C(i)\|^2$$

where $C(i)$ denotes the additional surveillance data on ith time interval. Assume τ' denotes the time interval between two consecutive data points of C, and τ is the interval of time step of the discrete simulation system. T' is defined as the number of timepoints of the surveillance data such that $T' = \lfloor T \cdot \tau'/\tau \rfloor$, $a_s = \lfloor i \cdot \tau'/\tau \rfloor$, $a_e = \lfloor (i+1) \cdot \tau'/\tau \rfloor - 1$. λ_2 is the scaling parameter.

6.7 Experiments

In this section, the performance of the proposed SimNest model is evaluated. First, the experiment setup is elaborated. Then, the effectiveness of the SimNest model on state-level influenza epidemic forecasting is demonstrated on real data by comparing with 8 comparison methods. In addition, the performance for forecasting fine-grained geographical subregion is evaluated.

6.7.1 Experiment Setup

This subsection presents the data preparation, label set and performance metrics.

Dataset

The Twitter data in this work was retrieved by the following process. First, the Twitter API is queried with flu-related keywords and the data during Jan 1, 2011 and Apr 15, 2015 in the United States is retrieved. The flu-related keywords include terms such as "flu", "influenza", and "h1n1", among others. The retrieved tweets are then classified according to whether or not they indicate the infection of their authors. The positive tweets are extracted and formed the influenza Twitter set, denoted as $\mathcal{D}_{(+)}$. For the classifier, LibShortText [107] is adopted, which is a logistic regression model specially designed for classifying short text like tweets. The classifier is trained on the existing labeled training set provided by Lamb et al. [62]. This training set forms the labeled tweets set, namely the tweets $\mathcal{X}_1$ and their labels $\mathcal{Y}_1$ in Section 6.4. The input features $\mathcal{K}$ of this model are the disease keywords provided by Paul and Dredze [81].

The authors $\mathcal{U}_2$ of the positive tweets set $\mathcal{D}_{(+)}$ are extracted and their tweets posted during two weeks before and after their tweets in $\mathcal{D}_{(+)}$ are retrieved via Twitter API. After removing

Table 6.1: Twitter data set and demographics

	Demographics		Twitter	
state	population size	#connections	#tweets	#users
CT	3,518,288	175,866,264	9,513,741	10,257
DC	599,657	19,984,180	12,148,925	7,015
MA	6,593,587	332,194,314	19,785,147	15,005
MD	5,699,478	285,159,648	20,754,218	19,758
VA	7,882,590	407,976,012	15,899,713	14,302

retweets, this Twitter data set is geocoded and only those tweets with location of interest are retained to form the unlabeled Twitter data set $\mathcal{X}_2$ defined in Section 6.4. Four states, including Connecticut (CT), Massachusetts (MA), Maryland (MD), and Virginia (VA), and the District of Columbia (DC) are utilized for this performance evaluation. The Carmen geocoder [38] is utilized to resolve the location of each tweet into a tuple containing information at the country, state, county, and city level. About 70% of the tweets in the dataset are assigned with a location by Carmen. To generate the contact network, the real demographics for each region is utilized. Substantial information about Twitter data and the demographics for the five regions are shown in Table 6.1.

Labels and Metrics

For the proposed model and all the competing methods, the data between Aug 1, 2011 and Jul 31, 2012 is utilized as the training season, while the data between Aug 1 2012 and Jul 31 2014 is used for predicting. The forecasting results for the flu outbreaks are validated against the corresponding influenza statistics reported by the Centers for Disease Control and Prevention (CDC). The CDC weekly publishes the percentage of the number of physician visits related to influenza-like illness (ILI) within each major region in the United States. In the experiment, four metrics are adopted, namely mean squared error (MSE), Pearson correlation, p-value, and peak time error. MSE stands for the mean value of the squared errors between all the predicted data points and corresponding label points. Pearson correlation is the covariance of the predicted and label data points divided by the product of their standard deviations. It varies from -1 to 1 and the larger the value, the stronger the positive correlation between them. The p-value denotes how likely the hypothesis of no correlation between the predicted and label data points is true. Thus, the smaller the p-value, the Pearson correlation is more statistically significant. Lastly, peak time error is the time interval between the predicted peak time (i.e., the week with the highest infectious number) and the actual peak time reflected by the CDC label data.

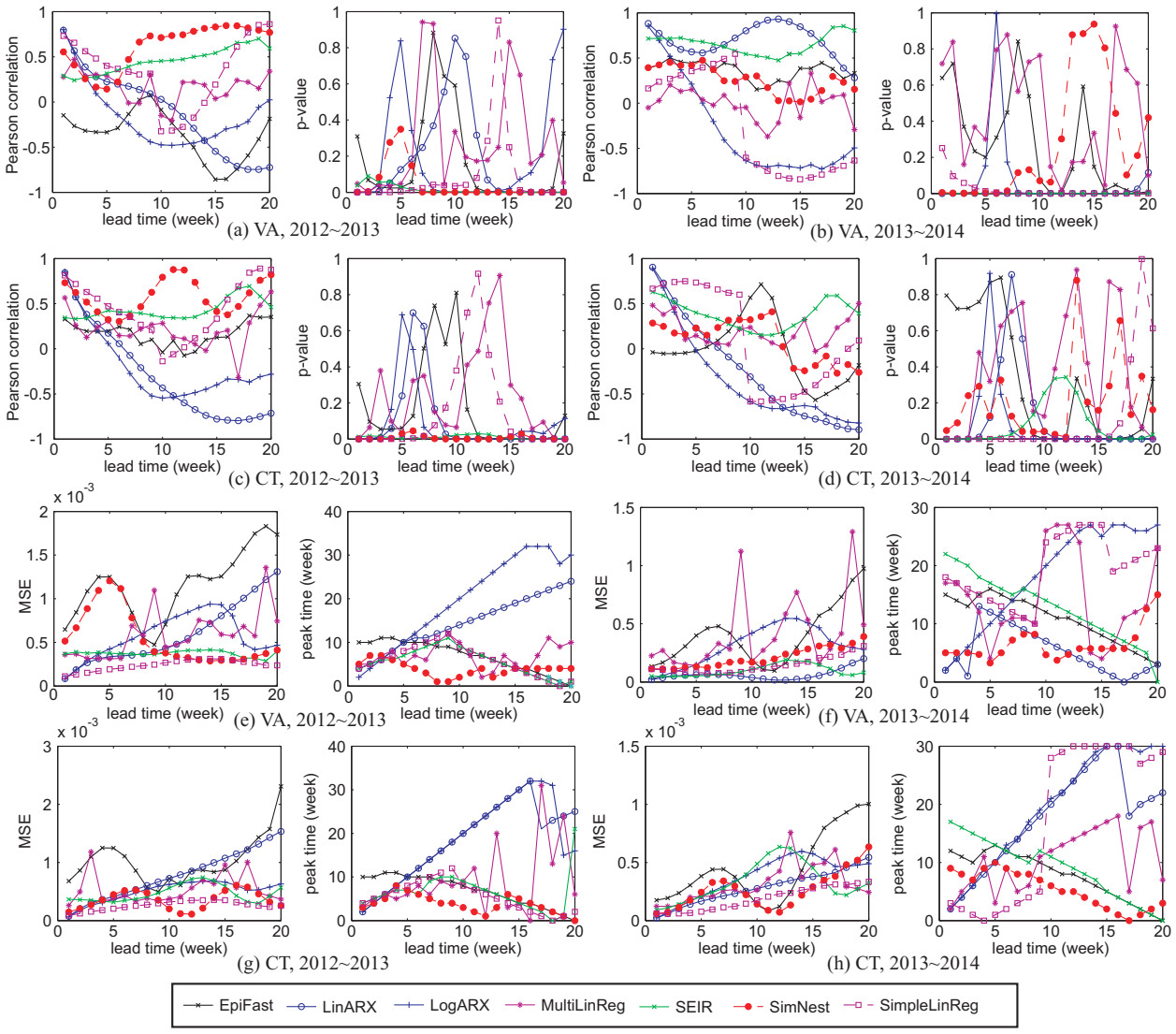


Figure 6.4: ILI visits percentage forecasting performance on the Pearson correlation and p-value for VA and CT in 3 seasons

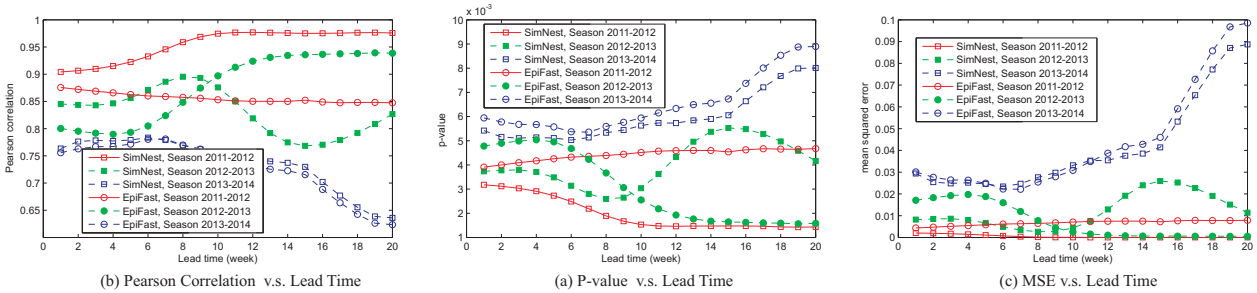


Figure 6.5: ILI visit percentage forecasting performance for Spatial subregions in CT for three flu seasons

6.7.2 State-level Influenza Epidemic Forecasting Performance

The performance for forecasting the percentage of ILI visits for each state with different lead times is evaluated. Specifically, the lead time vary from 1 week to 20 weeks, which means every method forecasts the data point from 1 week until 20 weeks in the future. Due to space limitations, only the results for Virginia and Connecticut are shown; the results for the other states exhibits the similar patterns to these two. In the experiment, the proposed SimNest model involves the extensions elaborated in Section 6.6.

Comparison methods

The proposed SimNest model is compared with 6 other methods. Among them, 4 methods are from social media mining: *Linear Autoregressive Exogenous model (LinARX)* [2], *Logistic Autoregressive Exogenous model (LogARX)* [3], *Simple Linear Regression model (simpleLinReg)* [51], *Multi-variable linear regression model (multiLinReg)* [37]. Another 2 methods are from computational epidemiology: *SEIR* [76] and *EpiFast* [14]. Their detailed settings are introduced as below. (1) *Linear Autoregressive Exogenous model (LinARX)* [2, 83]: This is a standard ARX model that builds the dependence of future visit percentage on the historical time series of CDC ILI visit percentage data [28] and volume of influenza tweets data $\mathcal{D}_{(+)}$. The orders of LinARX for Twitter data time series and CDC time series are set as 2 and 3 respectively based on cross-validation. (2) *Logistic Autoregressive Exogenous model (LogARX)* [3]: On the basis of LinARX, this method add a logit function transformation on the historical time series to enforce the boundary 0-1 of the value of ILI visit percentage. The orders of LogARX for the two time series are both set as 2 based on cross-validation. (3) *Simple Linear Regression model (simpleLinReg)* [51]: This method assumes a linear mapping between the input, the volume of infectious tweets $\mathcal{D}_{(+)}$, and the output, the future ILI visit percentages. (4) *Multi-variable linear regression model (multiLinReg)* [37]: This method treats a combination of keywords $\mathcal{K}$'s volumes as a multivariate input of the simple regression model. (5) *SEIR* [76]: This model divides the population into four health states, namely susceptible (S), exposed (E), infectious (I), and recovered (R). The epidemic dynamics are modeled by ordinary differential equations. The visit percentage is calculated through multiplying the volume of the state "I" by a ratio, which is optimized by cross-validation. (6) *EpiFast* [14]: This model follows the definition in Section 6.3, there are mainly two parameters that need to tune, p_E and p_I. They are optimized by minimizing the error of the predicted and the actual ILI visit percentage via Neald Mead method [18]. (7) *semi-supervised MLP+LinARX (semiLinARX)*: This method builds the classifier $f_W(\cdot)$ by simultaneously minimizing the loss functions $\mathcal{L}_1$, $\mathcal{L}_3$, and $\mathcal{L}_4$ in Equations 6.3, 6.8, and 6.9. The volume of the positive tweets classified is fed into LinARX. The orders of the LinARX for both time series of Twitter data CDC and surveillance data are set as 2 based on cross-validation. (8) *semi-supervised MLP+LogARX (semiLogARX)*: Using the same semi-supervised MLP as semiLinARX, the volume of the positive tweets classified is

fed into LogARX. The orders of LogARX for both time series of Twitter data and CDC surveillance data are set as 2 based on cross-validation.

Performance on the Pearson correlation and p-value

Figure 6.4(a), 6.4(b), 6.4(c), and 6.4(d) show the forecasting performance in terms of the Pearson correlation and p-value in two states, VA and CT, and for three seasons, 2011-2012, 2012-2013, and 2013-2014. Note that every season starts from August 1st and ends at July 31 each year. Also remember that the training period is 2011-2012 while the rest two seasons are both for testing. Overall, social media-based methods (i.e., LinARX, LogARX, MultiLinReg, and SimpleLinReg) typically achieves high Pearson correlation (i.e., between 0.6-0.95) with small lead time less than 2 weeks, but the Pearson correlation decreases all the way below 0 while lead time increases to 20. The p-value confirms the statistically significance of the high Pearson correlation when the lead time is less than 2 weeks. Computational epidemiology-based methods (i.e., SEIR and EpiFast), on the other hand, performs not as well as social media-based methods with small lead time, but the Pearson correlation does not drop significantly when lead time increases. For example, SEIR still can achieve a Pearson correlation around 0.6 while the lead time is 20 weeks. The reasons are two-folded. First, social media-based methods benefit from the real-time surveillance data while computational epidemiology-based methods use CDC data with a 1-2 week time lag. This difference makes the former one advantageous in predicting data points in the nearest future. Second, social media-based methods are purely data-driven, while computational epidemiology methods make use of the long-term disease progression mechanism. This makes computational epidemiology not too sensitive to current data and more robust in the performance. Among all the methods, the proposed SimNest model performs the best in overall performance by achieving the highest Pearson correlations in Figure 6.4(a), 6.4(c), and among the top 3 in Figure 6.4(b), and 6.4(d). In addition, the consistent low p-value indicates the robustness of the proposed SimNest model. This result demonstrates that SimNest successfully takes the advantages of the strengths of both social media-based methods and computational epidemiology-based methods.

Performance on MSE and peak time error

Figure 6.4(e), 6.4(f), 6.4(g), and 6.4(h) illustrate the performance on MSE and peak time error of all the methods in VA and CT for three seasons. Similar to the facts reflected by the Pearson correlation in Figure 6.4, the social media-based methods outperform computational epidemiology-based methods like SEIR and EpiFast in small lead time by achieving low MSE and peak time error. However, while the lead time increases, both the two errors of increase by 5-10 times. Computational epidemiology-based methods consistently achieves a reasonably well MSE and peak time error as low as 2-5 weeks. SimNest, again outperforms all the other methods in overall performance. Specifically, It achieves an MSE less than 5×10^{-4}

consistently in both training and testing periods, and achieves the peak time error around 0-4 weeks, which is generally 5-15 weeks less than that of social media-based methods, and at least 3-5 weeks less than that of computational epidemiology-based methods.

6.7.3 Spatial Subregion Outbreaks Forecasting Performance

Individual-based network epidemiology methods such as EpiFast can model the geographically detailed epidemic outbreaks. To demonstrate the advantage of embedding social media as an individual-level surveillance data, Figure 6.5 illustrates the comparison between the forecasting of ILI visit percentage for different subregions (i.e., counties) within the Connecticut state. According to Figure 6.5(a) and 6.5(b), The SimNest model outperforms EpiFast in the Pearson correlation for Season 2011-2012, Season 2013-2014, and half of Season 2012-2013. The p-values of both methods are less than 0.01 for all the three seasons, showing a statistically significance on the Pearson correlation comparison of them. Finally, the SimNest model again outperforms EpiFast in MSE for Season 2011-2012, Season 2013-2014, and half of Season 2012-2013.

6.8 Conclusions

To achieve timely and accurate epidemic diffusion modeling, computational epidemiology and social media mining communities recently have achieved important progress but still suffer from their different drawbacks. This work proposes SimNest, a novel bispace co-evolving framework to integrate the complementary strengths of computational epidemiology and social media mining. Extensive experiments based on multiple states and flu seasons demonstrated the advantages of integrating the respective strengths of computational epidemiology and social media mining. The detailed geographical subregion outbreaks forecasting is also improved by using social media that provides individual-level surveillance data.

Chapter 7

Conclusions and Future Work

The research presented here has focused on the development of spatiotemporal event detection and forecasting methods based on social media data. Four main types of approach were used, namely dynamic query expansion for event detection, generative framework for event forecasting, multi-task learning for spatiotemporal event forecasting, and deep learning based epidemics modeling for flu forecasting. To address the problem of dynamic query expansion for event detection, a novel unsupervised approach has been presented for detecting spatial events under targeted domains, along with a new dynamic query expansion that utilizes a Twitter heterogenous information network to dynamically extract domain-related key terms. To extract spatial events based on these domain-related tweets, a local modularity spatial scan capable of simultaneously considering the semantic similarity and the geographical proximities of tweets was designed. Extensive empirical studies on civil unrest event detection were conducted based on Twitter data collected in 10 Latin American countries. The results demonstrate the effectiveness and efficiency of the new approach. Future work will extend this method to cover 1) various types of topics, such as traffic events, crime events; and 2) various locations with different languages, such as in the Middle East.

To deal with the problem of creating a generative framework for event forecasting, a new generative approach was developed to uncover the underlying development of events by jointly considering the structural semantics and the spatiotemporal burstiness of Twitter streams. Both batch and online-based inference algorithms were developed to optimize the model parameters and then the trained model was utilized to calculate the alignment likelihood of tweet sequences via dynamic programming. Extensive empirical testing demonstrated the effectiveness of the new approach by comparing it with five representative methods. In the future, this approach can be extended to other applications, such as forecasting other disease outbreaks and local events such as road congestion. Further work will involve to the utilization of different distributions to model spatial burstiness.

With regard to the problem of multi-task learning for spatiotemporal event forecasting, this work presents a novel multi-task learning framework for spatial event forecasting in Social

Media. Existing methods are not able to concurrently address critical challenges such as the dynamic patterns of features, and geographic heterogeneity. This research treated the estimation of predictive models in different locations as a multi-task learning problem, in order to use shared information between locations, which effectively increases the sample size for each location. Both static and dynamic features were then modeled using different constraints to balance both homogeneity and diversity between these two types of features. The efficient algorithms proposed are based on the IGHT and are able to predict spatial events in real time. The empirical results presented here demonstrate that this method can be applied very effectively to detect civil unrest events, outperforming competing methods by a substantial margin on both precision and recall. In the future, the plan is to extend this multi-task learning framework by exploring more complex relationships between locations and integrating human domain knowledge as priors.

To address the problem of deep learning based epidemic modeling for flu forecasting and achieve timely and accurate epidemic diffusion modeling, the computational epidemiology and social media mining communities have begun to achieve important progress, although the approaches suggested still suffer from a number of different drawbacks. This work proposes SimNest, a novel bispace co-evolving framework that integrates the complementary strengths of computational epidemiology and social media mining. Extensive experiments based on multiple states and flu seasons demonstrated the advantages of integrating the respective strengths of computational epidemiology and social media mining. The accuracy of detailed geographical subregion outbreaks forecasting is also improved by using social media that provides individual-level surveillance data. Future work will be extended to new regions and new types of epidemics other than influenza outbreaks.

7.1 Contributions

The major research tasks performed are listed in Table 7.1, along with their current status.

7.1.1 Dynamic Query Expansion for Event Detection

- **Development of an unsupervised framework (A1)** A novel unsupervised approach for targeted domain spatial event detection in Twitter has been proposed. The new method requires no intensive human labor such as training set labeling.
- **Design of a novel dynamic query expansion (DQE) method (A2)** Given a targeted domain, DQE dynamically generates a set of domain-related key terms via a Twitter heterogeneous information network. The key terms are exhaustively extracted and then weighted appropriately based on DQE's iterative process.
- **An innovative local modularity spatial scan (LMSS) algorithm (A3)** Based on a

Table 7.1: Research tasks and status

Task	**Description**	**Status**
Research Area A	Dynamic Query Expansion	Completed
Research Area B	Generative Framework for Spatiotemporal Event Forecasting	
B1	Proposal of topic model based methods	Completed
B2	Proposal of parameter inference algorithm	Completed
B3	Extension to New Spatial-distribution modeling	Completed
B4	Extension to online parameter inference algorithm	Completed
B5	Extensions of the experiments	Completed
Research Area C	Multitask Learning for Spatiotemporal Event Forecasting	
C1	Proposal of multitask learning methods for event forecasting	Completed
C2	Proposal of parameter estimation algorithm	Completed
C3	Proposal of new model	Completed
C4	Validate performance on flu dataset	Completed
Research Area D	Deep learning based Epidemics modeling for Flu Forecasting	
D1	Proposal of Social media embedded epidemics model	Completed
D2	Extension to Dynamics of social network.	Completed
D3	Extension to Heterogeneous Surveillance Data.	Completed
D4	Extend experiment to new datasets and baselines.	Completed
Research Area E	Event Forecasting on Hierarchical-multisource Indicators	
E1	Prepare data set, label set, and baseline methods.	Completed
E2	Use ADMM to solve our 2-level interactive lasso problem and evaluate the performance.	Completed
E3	Solve our 3-level interactive lasso problem.	Completed
E4	Handle incomplete data by multitask learning	Completed
Thesis Revision F	Thesis Revision	Completed

graph formed using key terms from DQE, LMSS jointly maximizes the local modularity and spatial scan statistics in order to distinguish events by taking into account both their semantic similarities and geographical proximities.

7.1.2 Generative Framework for Spatiotemporal Event Forecasting

- **A novel generative framework for spatial event forecasting (B1)** For spatial event forecasting in Twitter, an enhanced hidden Markov model (HMM) has been proposed that characterizes the transitional process of event development by jointly considering the time-evolving context and space-time burstiness of Twitter streams.
- **Effective batch and online algorithms for model parameter inference (B2)** The model inference is formalized as the maximization of a posterior that is analytically tractable. Both EM-based algorithm and stochastic-EM parameter optimization algorithms are proposed to solve this problem effectively and efficiently.
- **Extension to New Spatial-distribution modeling (B3)** In addition to the original model presented at the conference SDM 2015, another new model has been proposed that is based on the use of a Gamma-Poisson distribution to handle positive values of spatiotemporal burstiness.
- **Extension to online parameter inference algorithm (B4)** New algorithms have been proposed that is build on the work presented in a previous SDM paper, which considered only a batch-based parameter inference algorithm. Of the two new online algorithms for parameter inference, one is for the original model and the other for the newly proposed Gamma-Poisson distribution-based model.
- **Extensions of the experiments (B5)** Scalability studies that incorporate scalability validation and analysis for both models have been added to the two different algorithms introduced above and experiments conducted on two different datasets, namely the flu dataset and the civil unrest dataset. The results clearly demonstrated that the newly-proposed online algorithm is much more scalable than the batch-based algorithm and is also more efficient, especially on large datasets. Case studies for civil unrest event forecasting have been presented and extensively analyzed. All the important components of the new model, including the spatial burstiness for all the latent states, the distribution of words for all the latent topics, and the distribution of topics for all the latent states, have been illustrated and examined. Finally, the forecasting of a specific civil unrest event has been demonstrated. Both the transitions among the latent states and the discovery of event-specific keywords have been illustrated and validated against the ground-truth, which in this case consist of news reports covering this event from authorized news outlets.

7.1.3 Multitask Learning for Spatiotemporal Event Forecasting

- **Formulation of a multi-task learning framework for event forecasting (C1)** Event forecasting for multiple cities in the same country can be formulated as a multi-task learning problem. In the proposed model, event forecasting models were built for different cities simultaneously by restricting all cities to select a common set of features and both penalized and constrained MTL formulations, which use different strategies to control the common set of features selected, explored.
- **Concurrent modeling of static and dynamic terms (C2)** The existing models (LASSO and DQE) use different but complementary information; LASSO uses static terms, while DQE identifies dynamic terms. The new MTL formulations make use of both types of information by integrating the strengths of LASSO (a supervised approach) and DQE (an unsupervised approach). To the best of our knowledge, there is little prior work that combines supervised and unsupervised approaches for event forecasting.
- **Proposal of a new model (C3)** In addition to the proposed multitask model, a new model has been designed to constrain either the static or dynamic features. Specifically, this model enables users to specify the intended number of selected dynamic features and then determines the number of selected static features automatically.
- **Extend the experimental evaluation of the flu dataset (C4)** A new dataset has been added in order to evaluate the current methods and use CDC surveillance data as the label set. Sensitivity analysis of the initial W of the hard thresholding has been added, along with a scalability analysis, and tested experimentally.

7.1.4 Deep Learning Based Epidemic Modeling for Flu Forecasting

- **A novel integrated framework for computational epidemiology and social media mining (D1)**: Existing approaches from computational epidemiology and social media mining focus on different but complementary aspects. The former focuses on modeling the underlying mechanisms of disease diffusion while the latter provides timely and detailed disease surveillance. The new SimNest framework utilizes both types of information by integrating their respective strengths.
- **A semi-supervised multilayer perceptron (MLP) for mining epidemic features (D2)**: To achieve deep integration, unsupervised pattern constraints are enforced that have been derived from an epidemic disease progress model onto the supervised classification. Using this semi-supervised strategy, the sparsity of labeled data can be solved.
- **Extension to the dynamics of social networks (D3)** As the epidemic diffusion process progresses, interventions are among the most common and effective ways for the

government and individuals to reduce the potential impact of disease outbreaks. Interventions influence the epidemic diffusion process largely by changing the people-people contact network. These interventions can be categorized into two types: (1) Pharmaceutical (PI) and (2) Non-pharmaceutical (NPI). PI interventions, such as administering antivirals and vaccines, can change the characteristics (e.g., disease transmissibility) of the person nodes in the social contact network, while NPI interventions are those actions that effectively change the contact network structure, including school closures, quarantine and sequestration. Therefore, both types of interventions can result in changes in the social contact network.

- **Utilization of the Heterogeneous Surveillance Data (D4)** The SimNest framework is sufficiently flexible to incorporate multiple surveillance data sources. In the basic problem definition, social media data is used solely to provide fine-grained surveillance data but SimNest makes it possible to include heterogeneous surveillance data sources such as CDC [28] surveillance data for the United States, and Paho [80] surveillance data for Latin America.

- **Evaluations on new datasets and baselines (D5)** Adding new datasets makes it possible to evaluate the current methods more rigorously and include more baselines for comparison. In addition, disease epidemics other than influenza could be considered in future research.

7.1.5 Event Forecasting on Hierarchical Multisource Indicators

- **Prepare data set, label set, and baseline methods (E1)** This proposed research focuses on hierarchical-structure-based multisource inputs from social media data. The work of data integration and preprocessing is nontrivial and the development of new baseline methods for this complex problem also requires a considerable amount of effort.

- **Use ADMM to solve the 2-level interactive Lasso problem and evaluate the new model's performance (E2)** Hierarchical-structure-based multisource inputs can be used to formulated a hierarchical linear regression problem by considering country-level and city-level inputs. This will allow group Lasso to be utilized to regularize the hierarchical feature selection process under the sparsity assumption. Rigorous experimental evaluations are then conducted.

- **Solve the 3-level interactive Lasso problem (E3)** Beyond 2-level hierarchical features, more complex 3-level features are considered, including country-level, state-level, and city-level features. Similar techniques are applied, although it will be necessary to formulate brand new objective functions due to the involvement of new interactive feature terms.

- **Handle incomplete data via multitask learning (E4)** When many data sources are involved, it is unrealistic to assume that all the data sources have the same time range. In order to utilize all the overlapping time range and the non-overlapping time range effectively across all the data sources, multitask learning is implemented to handle temporally incomplete multisource data.

7.2 Publications

7.2.1 Published Papers at VT

Journal papers

1. Liang Zhao, Feng Chen, Jing Dai, Ting Hua, Chang-Tien Lu, and Naren Ramakrishnan. "Unsupervised Spatial Event Detection in Targeted Domains with Applications to Civil Unrest Modeling." PLoS ONE (impact factor: 3.534), vol. 9, no. 10 (2014): e110206.

2. Liang Zhao, Ting Hua, Chang-Tien Lu, and Ing-Ray Chen. "A Topic-focused Trust Model for Twitter." Computer Communications, (impact factor: 1.7), Elsevier, vol. 76, pp. 1-11, Feb 2016.

3. Jinliang Ding, Liang Zhao, Changxin Liu, and Tianyou Chai. "GA-based principal component selection for production performance estimation in mineral processing." Computers and Electrical Engineering, vol. 40, no. 5 (2014): 1447-1459.

4. Ting Hua, Feng Chen, Liang Zhao, Chang-Tien Lu and Naren Ramakrishnan. Automatic Targeted-Domain Spatiotemporal Event Detection in Twitter. Geoinformatica, accepted.

5. Fang Jin, Wei Wang, Liang Zhao, Edward Dougherty, Yang Cao, Chang-Tien Lu, and Naren Ramakrishnan. "Misinformation Propagation in the Age of Twitter." IEEE Computer, vol. 47, no. 12 (2014): 90-94.

6. Andy Doyle, Graham Katz, Kristen Summers, Chris Ackermann, Ilya Zavorin, Zunsik Lim, Sathappan Muthiah, Liang Zhao, Chang-Tien Lu, Patrick Butler, Rupinder Paul Khandpur. "Forecasting Significant Societal Events Using The Embers Streaming Predictive Analytics System." Big Data Journal, vol. 2, no. 4 (2014): 185-195.

Conference papers

1. Liang Zhao, Jieping Ye, Feng Chen, Chang-Tien Lu, Naren Ramakrishnan. Hierarchical Incomplete Multisource Feature Learning for Spatiotemporal Event Forecasting. in

Proceedings of the 22st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2016), research track, accepted (acceptance rate: 18.2%).

2. Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. "Multi-Task Learning for Spatio-Temporal Event Forecasting." in Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2015), research track, (acceptance rate: 19.4%), Sydney, Australia, pp. 1503-1512, Aug 2015.

3. Liang Zhao, Jiangzhuo Chen, Feng Chen, Wei Wang, Chang-Tien Lu, and Naren Ramakrishnan. "SimNest: Social Media Nested Epidemic Simulation via Online Semi-supervised Deep Learning." in Proceedings of the IEEE International Conference on Data Mining (ICDM 2015), regular paper (acceptance rate: 8.4%), Atlantic City, NJ, pp. 639-648, Nov 2015.

4. Liang Zhao, Feng Chen, Chang-Tien Lu, Naren Ramakrishnan. "Dynamic Theme Tracking in Twitter." in Proceedings of the IEEE International Conference on Big Data (BigData 2015), regular paper (acceptance rate: 16.8%), Santa Clara, California, pp. 561-570, Oct-Nov 2015.

5. Liang Zhao, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. "Spatiotemporal Event Forecasting in Social Media." in Proceedings of the SIAM International Conference on Data Mining (SDM 2015), (acceptance rate: 22%), Vancouver, BC, pp. 963-971, Apr-May 2015.

6. Sathappan Muthiah, Patrick Butler, Rupinder Paul Khandpur, Parang Saraf, Nathan Self, Alla Rozovskaya, Liang Zhao, Jose Cadena, Chang-Tien Lu, Anil Vullikanti, Achla Marathe, Kristen Summers, Graham Katz, Andy Doyle, Jaime Arredondo, Dipak K. Guptak, David Mares, Naren Ramakrishnan. EMBERS at 4 years: Experiences operating an Open Source Indicators Forecasting System. in Proceedings of the 22st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2016), applied data science track, (acceptance rate: 19.9%), San Francisco, California, Aug 2016.

7. Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, Chris Kuhlman, Achla Marathe, Liang Zhao, 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 (KDD 2014), industrial track, pp. 1799-1808. ACM, 2014.

8. Hua, Ting, Feng Chen, Liang Zhao, Chang-Tien Lu, and Naren Ramakrishnan. "STED: semi-supervised targeted-interest event detectionin in twitter." InProceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2013), demo track, pp. 1466-1469. ACM, 2013.

9. Andy Doyle, Graham Katz, Kristen Summers, Chris Ackermann, Ilya Zavorin, Zunsik Lim, Sathappan Muthiah, Liang Zhao, Chang-Tien Lu, Patrick Butler, Rupinder Paul Khandpur. "The EMBERS architecture for streaming predictive analytics." In Proceedings of the IEEE International Conference on Big Data (BigData 2014), pp. 11-13. IEEE, 2014.

7.2.2 Submitted Journal Papers

1. Liang Zhao, Feng Chen, Chang-Tien Lu, Naren Ramakrishnan. Online Spatial Event Forecasting in Microblogs. ACM Transactions on Spatial Algorithms and Systems (TSAS), in second-round review.

2. Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Feature Constrained Multi-Task Learnings for Event Forecasting in Social Media." IEEE Transactions on Knowledge and Data Engineering (TKDE), major revision.

3. Liang Zhao, Jiangzhuo Chen, Feng Chen, Fang Jin, Wei Wang, Chang-Tien Lu, Naren Ramakrishnan. " Social Media-driven Online Epidemics Modeling by Adaptive Semi-supervised Multilayer Perceptron", ACM Transactions on Knowledge Discovery from Data (TKDD), submitted.

4. Yan Shi, Min Deng, Qiliang Liu, Liang Zhao, Chang-Tien Lu. A framework to discover domain related spatio-temporal evolving patterns from Twitter. International Journal of Geo-Information, submitted.

7.2.3 Submitted Conference Papers

1. Liang Zhao, Feng Chen, Chang-Tien Lu, Naren Ramakrishnan. Multi-resolution spatial event forecasting in social media. in International Conference on Data Mining (ICDM 2016), submitted.

2. Xuchao Zhang, Liang Zhao, Zhiqian Chen, Arnold P. Boedihardjo, Dai Jing, and Chang-Tien Lu. Trendi: Tracking Stories in News and Microblogs via Emerging, Evolving and Fading Topics. in International Conference in Data Mining (ICDM 2016), submitted.

3. Rupinder Paul Khandpur, Taoran Ji, Yue Ning, Liang Zhao, Chang-Tien Lu, Erik R. Smith, Christopher Adams, Naren Ramakrishnan. Determining Relative Airport Threats from News and Social Media. The 25th ACM International Conference on Information and Knowledge Management (CIKM 2016), submitted.

7.3 Future Research Directions

7.3.1 Dynamic Keyword Expansion in Social Media

1) Consideration of Languages heterogeneity. Multiple language postings are very common in social media in the regions like the United States and Latin America. We plan to embed transfer learning to our dynamic query expansion procedures to strengthen the learning process for respective languages.

2) Extensions to more regions and domains. Because our method is unsupervised which is cheap to redistribute and effective to handle dynamic nature of social media data, we are working on extending current method to applications for MENA region. We are also working on the new applications of event detection on transportation, crimes, and climate.

7.3.2 Spatiotemporal Event Forecasting in Social Media

1) Multi-indicator-based predictive modeling. The multi-indicator-based predictive modeling can handle the prediction of complex social events caused by multiple social factors comprehensively. The challenges including hierarchical data structure, incomplete data, and multiple spatial resolution will be simultaneously handled.

2) Graph-structured multitask learning. Graph-structured multitask learning is an extension of our multitask learning framework in [112]. In the new work, we will quantify the strengths of tasks correlations in multiple aspects by a new heterogeneous-graph-structured multitask learning framework.

3) Event-subtype forecasting. Other than only predicting the occurrence of future events, we plan to also forecast the subtypes and topics of the future event based on higher-order multitask learning framework, where the tasks are defined not in spatial dimension but also in subtype dimension.

7.3.3 Social Media-embedded Influenza Epidemics Modeling

1) Deeply mining the social media knowledge. We will identify in real time the intervention and inpatient actions from social media and incorporate them to adjust the corresponding parameters of epidemics simulation model.

2) Developing a distributed model. Jointly mining social media and simulating epidemics is a time consuming iterative process. Thus, we plan to propose new scalable approaches based on distributed framework.

Appendix A

Online Spatial Event Forecasting on Microblog Streams

A.1 Batch Parameter Optimization Algorithm

To find the best parameters for both models, the Expectation Maximization (EM) algorithm is extended to compute the parameters of the structural text and space-time outbreaks modeling.

The standard steps of the Baum-Welch (BW) algorithm [31] are applied to calculate the expectation probability $\mathrm{E}[p(Z_{s,t} = k)]$ that the observation of the location s and time t is under the latent states k.

Given the expectations $\mathrm{E}[p(Z_{s,t} = k)]$, the expected count of time intervals for the observations under the latent state k is calculated based on the following equations.

$$\hat{N}_{l,k} = \sum\nolimits_{s \subseteq D_l} \sum\nolimits_{t} \mathrm{E}[p(Z_{s,t} = k)] \tag{A.1}$$

where $s \in D_l$ denotes that the sequence s belongs to the tweets in the location l.

When using the Gaussian distribution to model the space-time burstiness, the maximum likelihoods of the mean and variance of the Gaussian distributed space-time burstiness modeling are computed as below:

$$\hat{\mu}_{l,k} = \frac{\sum_{s \subseteq D_l} \sum_{t} \mathrm{E}[p(Z_{s,t} = k)] \cdot (r^{in}_{s,t}, r^{out}_{s,t})}{\hat{N}_{l,k}} \tag{A.2}$$

where $(r^{in}_{s,t}, r^{out}_{s,t})$ is the vector observation of the bi-variate Gaussian.

$$\hat{\Sigma}_{l,k} = \frac{\sum_{s \subseteq D_l} \sum_{t} \mathrm{E}[p(Z_{s,t} = k)] \big(\hat{\mu}_{l,k} - (r^{in}_{s,t}, r^{out}_{s,t})\big)^2}{\hat{N}_{l,k}} \tag{A.3}$$

The posteriors of the mean and variance of the Gaussian distributed space-time burstiness modeling are computed as below:

$$\mu_{l,k} = (\beta_0\mu_0 + \hat{N}_{l,k}\hat{\mu}_{l,k}) \Big/ (\beta_0 + \hat{N}_{l,k}) \tag{A.4}$$

$$\Sigma_{l,k} = \frac{\Lambda_0 + \hat{\Sigma}_{l,k}}{\nu_0 + 3} + \frac{\beta_0\hat{N}_{l,k}(\hat{\mu}_{l,k} - \mu_0)(\hat{\mu}_{l,k} - \mu_0)^T}{(\beta_0 + \hat{N}_{l,k})(\nu_0 + 3)} \tag{A.5}$$

When using Poisson-distributed space-time burstiness modeling, Equations A.2 $\sim$ A.5 are replaced by Equations A.6 and A.11, as shown in the following:

The weighted means of the domain-related counts inside location l under latent state k are calculated as below:

$$\hat{\lambda}^{in}_{c,k,l} = \sum\nolimits_{s\in D_l}\sum\nolimits_t c^{in}_{s,t} \cdot \mathrm{E}[p(Z_{s,t} = k)]/\hat{N}_{l,k} \tag{A.6}$$

The weighted means of the base counts inside location l under latent state k are calculated as below:

$$\hat{\lambda}^{in}_{b,k,l} = \sum\nolimits_{s\in D_l}\sum\nolimits_t b^{in}_{s,t} \cdot \mathrm{E}[p(Z_{s,t} = k)]/\hat{N}_{l,k} \tag{A.7}$$

The weighted means of the domain-related counts outside location l under latent state k are calculated as below:

$$\hat{\lambda}^{out}_{c,k,l} = \sum\nolimits_{s\in D_l}\sum\nolimits_t c^{out}_{s,t} \cdot \mathrm{E}[p(Z_{s,t} = k)]/\hat{N}_{l,k} \tag{A.8}$$

The weighted means of the base counts outside location l under latent state k are calculated as below:

$$\hat{\lambda}^{out}_{b,k,l} = \sum\nolimits_{s\in D_l}\sum\nolimits_t b^{out}_{s,t} \cdot \mathrm{E}[p(Z_{s,t} = k)]/\hat{N}_{l,k} \tag{A.9}$$

Therefore, the posteriors of the means of the ratios of the domain-related tweets inside and outside location l under the latent state k are calculated as the following two equations, respectively:

$$\lambda^{in}_{k,l} = \frac{(\alpha^{in} - 1) + \hat{\lambda}^{in}_{c,k,l}}{\beta^{in} + \hat{\lambda}^{in}_{b,k,l}} \tag{A.10}$$

$$\lambda^{out}_{k,l} = \frac{(\alpha^{out} - 1) + \hat{\lambda}^{out}_{c,k,l}}{\beta^{out} + \hat{\lambda}^{out}_{b,k,l}} \tag{A.11}$$

In sequence s and latent state k, the expectations of the count of the word w that is labeled as being a word that is specific to the unique event is calculated as below:

$$g_{s,k,w} = \sum_t^T N_{s,t,w}\frac{\mathrm{E}[p(Z_{s,t} = k)] \cdot \Psi_{k,1} \cdot \theta^R_{s,t,w}}{\Psi_{k,1}\theta^R_{s,t,w} + \Psi_{k,2}\sum_j \Phi_{k,j}\theta^{B_j}_w} \tag{A.12}$$

In latent state k, the expectation of the count of the word w that is labeled as being a common word under latent topic j is calculated as below:

$$f_{k,j,w} = \sum_{s}^{S} \sum_{t}^{T} N_{s,t,w} \frac{\mathrm{E}[p(Z_{s,t} = k)] \cdot \Psi_{k,2} \cdot \Phi_{k,j} \theta_w^{B_j}}{\Psi_{k,1} \theta_{s,t,w}^{R} + \Psi_{k,2} \sum_j \Phi_{k,j} \theta_w^{B_j}} \tag{A.13}$$

Among the words corresponding to specific events in sequence s and latent state k, the likelihood of the word w occurring is:

$$\theta_{s,k,w}^{R} = \frac{g_{s,k,w}}{\sum_x g_{s,k,x}} \tag{A.14}$$

Among the common words under topic j in sequence s and latent state k, the likelihood of the word w occurring is:

$$\theta_w^{B_j} = \frac{\sum_k f_{k,j,w}}{\sum_k \sum_w f_{k,j,w}} \tag{A.15}$$

In the latent state k, the likelihood that a word will correspond to a specific event is:

$$\Psi_{k,1} = \frac{\sum_s \sum_w g_{s,k,w}}{\sum_s \sum_w g_{s,k,w} + \sum_w \sum_j f_{k,j,w}} \tag{A.16}$$

In the latent state k, the likelihood that a word will be a common word is:

$$\Psi_{k,2} = \frac{\sum_w \sum_j f_{k,j,w}}{\sum_s \sum_w g_{s,k,w} + \sum_w \sum_j f_{k,j,w}} \tag{A.17}$$

In latent state k among all the common words, the likelihood that a word will fall under topic j is:

$$\Phi_{k,j} = \frac{\sum_w f_{k,j,w}}{\sum_j \sum_w f_{k,j,w}} \tag{A.18}$$

The prior likelihood of latent state k is:

$$\pi_k = \frac{\sum_s \mathrm{E}[p(Z_{s,1} = k)]}{\sum_s \sum_i \mathrm{E}[p(Z_{s,1} = i)]} \tag{A.19}$$

By iteratively executing the E-step and the M-step, the model parameters and the latent variables are continuously updated until convergence is achieved. The model parameters are optimized while the likelihood in Equation 3.10 is maximized.

A.2 Stochastic E-Step

A.2.1 STM-I

$E_i^{\hat{\mu}}$, $E_i^{\hat{\Sigma}}$, E_i^g, and E_i^f can be obtained based on the current sequence s_i as shown in Equations A.20.

$$\begin{aligned}
E_i^{\hat{\mu}} &= \frac{\sum_t \mathrm{E}[p(Z_{s_i,t}=k)](r_{s_i,t}^{in}, r_{s_i,t}^{out})}{\hat{N}_{l,k,i}} \\
E_i^{\hat{\Sigma}} &= \frac{\sum_t \mathrm{E}[p(Z_{s_i,t}=k)]\big(\hat{\mu}_{l,k} - (r_{s_i,t}^{in}, r_{s_i,t}^{out})\big)^2}{\hat{N}_{l,k,i}} \\
E_i^g &= \sum_t^T N_{s_i,t,w} \frac{\mathrm{E}[p(Z_{s_i,t}=k)] \cdot \Psi_{k,1} \cdot \theta_{s_i,t,w}^R}{\Psi_{k,1}\theta_{s_i,t,w}^R + \Psi_{k,2}\sum_j \Phi_{k,j}\theta_w^{B_j}} \\
E_i^f &= \sum_t^T N_{s_i,t,w} \frac{\mathrm{E}[p(Z_{s_i,t}=k)] \cdot \Psi_{k,2}\sum_j \Phi_{k,j}\theta_w^{B_j}}{\Psi_{k,1}\theta_{s_i,t,w}^R + \Psi_{k,2}\sum_j \Phi_{k,j}\theta_w^{B_j}}
\end{aligned} \tag{A.20}$$

The stochastic approximation of the statistics update is presented in Equations A.24

$$\begin{aligned}
\hat{\mu}_{l,k,i} &= (1-\gamma_i)\cdot\hat{\mu}_{l,k,i-1} + \gamma_i \cdot E_i^{\hat{\mu}} \\
\hat{\Sigma}_{l,k,i} &= (1-\gamma_i)\cdot\hat{\Sigma}_{l,k,i-1} + \gamma_i \cdot E_i^{\hat{\Sigma}} \\
g_{s,k,w,i} &= (1-\gamma_i)\cdot g_{s,k,w,i-1} + \gamma_i \cdot E_i^g \\
f_{k,j,w,i} &= (1-\gamma_i)\cdot f_{k,j,w,i-1} + \gamma_i \cdot E_i^g
\end{aligned} \tag{A.21}$$

The model parameters θ^B, θ^R, Ψ, and Φ need to be initialized. For each latent state $k \in K$, the common-word language model $\theta^B \in \mathbb{R}^{K\times J\times N}$ is initialized by maximizing the likelihood of a mixture multinomial model. Specifically,

$$p(w) = \sum_j^J p(j) \prod_n^N p(W_n|j) \tag{A.22}$$

where $p(W_n|j) = \theta_{n,k}^{B_j}$. By maximizing the log likelihood of $p(w)$, the language model θ^B is determined. Other parameters θ^R, Ψ, and Φ are initialized with uniform distributions.

A.2.2 STM-S

Specifically, the conditional expectations $E_i^{\hat{\mu}}$, $E_i^{\hat{\Sigma}}$, E_i^g, and E_i^f based on the currency sequence s_i are calculated.

$$\begin{aligned}
E_i^{\hat{\lambda}_c^{in}} &= \sum\nolimits_t c_{s_i,t}^{in} \cdot \mathrm{E}[p(Z_{s_i,t}=k)]/\hat{N}_{l,k,i} \\
E_i^{\hat{\lambda}_b^{in}} &= \sum\nolimits_t b_{s_i,t}^{in} \cdot \mathrm{E}[p(Z_{s_i,t}=k)]/\hat{N}_{l,k,i} \\
E_i^{\hat{\lambda}_c^{out}} &= \sum\nolimits_t c_{s_i,t}^{out} \cdot \mathrm{E}[p(Z_{s_i,t}=k)]/\hat{N}_{l,k,i} \\
E_i^{\hat{\lambda}_b^{out}} &= \sum\nolimits_t b_{s_i,t}^{out} \cdot \mathrm{E}[p(Z_{s_i,t}=k)]/\hat{N}_{l,k,i} \\
E_i^g &= \sum_t^T N_{s_i,t,w} \frac{\mathrm{E}[p(Z_{s_i,t}=k)] \cdot \Psi_{k,1} \cdot \theta_{s_i,t,w}^R}{\Psi_{k,1}\theta_{s_i,t,w}^R + \Psi_{k,2}\sum_j \Phi_{k,j}\theta_w^{B_j}} \\
E_i^f &= \sum_t^T N_{s_i,t,w} \frac{\mathrm{E}[p(Z_{s_i,t}=k)] \cdot \Psi_{k,2}\sum_j \Phi_{k,j}\theta_w^{B_j}}{\Psi_{k,1}\theta_{s_i,t,w}^R + \Psi_{k,2}\sum_j \Phi_{k,j}\theta_w^{B_j}}
\end{aligned} \tag{A.23}$$

The stochastic approximation of the statistics update is presented in Equations A.24.

$$\begin{aligned}
\hat{\lambda}_{c,l,k,i}^{in} &= (1-\gamma_i)\cdot\hat{\lambda}_{c,l,k,i-1}^{in} + \gamma_i \cdot E_i^{\hat{\lambda}_c^{in}} \\
\hat{\lambda}_{b,l,k,i}^{in} &= (1-\gamma_i)\cdot\hat{\lambda}_{b,l,k,i-1}^{in} + \gamma_i \cdot E_i^{\hat{\lambda}_b^{in}} \\
\hat{\lambda}_{c,l,k,i}^{out} &= (1-\gamma_i)\cdot\hat{\lambda}_{c,l,k,i-1}^{out} + \gamma_i \cdot E_i^{\hat{\lambda}_c^{out}} \\
\hat{\lambda}_{b,l,k,i}^{out} &= (1-\gamma_i)\cdot\hat{\lambda}_{b,l,k,i-1}^{out} + \gamma_i \cdot E_i^{\hat{\lambda}_b^{out}} \\
g_{s,k,w,i} &= (1-\gamma_i)\cdot g_{s,k,w,i-1} + \gamma_i \cdot E_i^g \\
f_{k,j,w,i} &= (1-\gamma_i)\cdot f_{k,j,w,i-1} + \gamma_i \cdot E_i^g
\end{aligned} \tag{A.24}$$

Appendix B

Deep Learning Based Epidemics Modeling for Flu Forecasting

B.1 The Derivatives with Respect to W

In this section, the partial derivatives of the loss function (in Equation 2 in the original paper) are elaborated with respect to the weight matrix W. This can be decomposed into the partial derivatives of each of the sub-loss functions $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$, and $\mathcal{L}_4$ in Equations 4, 6, 8, and 9, respectively.

$$\frac{\partial \mathcal{L}_{1,u,t}}{\partial W^{(1)}_{j,k}} = (f_W(X_{u,t}) - Y_{u,t}) s'(h^{(1)}) W^{(2)}_j s'(h^{(2)}_j) X^{(l)}_{i,k} \tag{B.1}$$

where $s'(x) = s(x) \cdot (1 - s(x))$.

$$\frac{\partial \mathcal{L}_{1,u,t}}{\partial W^{(2)}_{j}} = (f_W(X_{u,t}) - Y_{u,t}) \cdot s'(h^{(1)}) \cdot s(h^{(2)}_j) \tag{B.2}$$

where $\mathcal{L}_{1,u,t} = \mathcal{L}_1(f_W(X_{u,t}), Y_{u,t})$.

$$\begin{aligned}\frac{\partial \mathcal{L}_{4,u}}{\partial W^{(1)}_{j,k}} =& f_W(X_{u,t}) s'(h^{(1)}) W^{(2)}_j s'(h^{(2)}_j) X^{(n)}_{u,t,k} \\ & - f_W(X_{u,t+1}) s'(h^{(1)}) W^{(2)}_j s'(h^{(2)}_j) X^{(n)}_{u,t,k}\end{aligned} \tag{B.3}$$

where $\mathcal{L}_{4,u} = \sum_t^{\mathcal{T}} \mathcal{L}_4(X_{u,t}, X_{u,t+1}, W)$.

$$\frac{\partial \mathcal{L}_{4,u}}{\partial W^{(2)}_{j}} = (f_W(X_{u,t}) - f_W(X_{u,t+1})) s'(h^{(1)}) s(h^{(2)}_j) \tag{B.4}$$

The derivative of $\mathcal{L}_3$ with respect to W is deduced as follows:

$$\frac{\partial \mathcal{L}_{3,u}}{\partial W_{j,k}^{(1)}} = \sum_{t}^{T} \frac{\partial \mathcal{L}_{3,u,t}}{\partial s(h_t^{(1)})} \frac{\partial s(h_t^{(1)})}{\partial h_t^{(1)}} \frac{\partial h_t^{(1)}}{\partial s(h_t^{(2)})} \frac{\partial s(h_t^{(2)})}{\partial h_t^{(2)}} \frac{\partial h_t^{(2)}}{\partial W_{j,k}^{(1)}}$$
$$= \sum_{t}^{T} (\sum_{i}^{T} f_W(X_{u,i}) - p_I) s'(h_t^{(1)}) W_j^{(2)} s'(h_{t,j}^{(2)}) X_{u,t,k}^{(n)} \tag{B.5}$$

where $\mathcal{L}_{3,u} = \sum_t^{\mathcal{T}} \mathcal{L}_{3,u,t}$, and $\mathcal{L}_{3,u,t} = \mathcal{L}_3(X_{u,t}, W, p_I)$.

$$\frac{\partial \mathcal{L}_{3,u}}{\partial W_j^{(2)}} = \sum\nolimits_t^{\mathcal{T}} \frac{\partial \mathcal{L}_{3,u,t}}{\partial s(h_t^{(1)})} \frac{\partial s(h_t^{(1)})}{\partial h_t^{(1)}} \frac{\partial h_t^{(1)}}{\partial W_j^{(2)}}$$
$$= \sum_{t}^{T} (\sum_{i}^{T} f_W(X_{u,i}) - p_I) s'(h_t^{(1)}) s(h_{t,j}^{(2)}) \tag{B.6}$$

Similarly, the derivative of $\mathcal{L}_2$ is as follows:

$$\frac{\partial \mathcal{L}_{2,l,t}}{\partial W_{j,k}^{(1)}} = \sum_{u}^{\mathcal{U}_{2,l,t}} (\sum_{v}^{\mathcal{U}_{2,l,t}} f_W(X_{v,t}) - \sum_{v}^{\mathcal{V}_{l,t}} Q_v(p_E, p_I))$$
$$\cdot s'(h_u^{(1)}) W_j^{(2)} \cdot s'(h_{u,j}^{(2)}) \cdot X_{u,k} \tag{B.7}$$

where $\mathcal{L}_{2,l,t} = \mathcal{L}_2(X_{l,t}, W, \mathcal{Z})$.

$$\frac{\partial \mathcal{L}_{2,l,t}}{\partial W_j^{(2)}} = \sum_{u}^{\mathcal{U}_{2,l,t}} \sum_{v}^{\mathcal{U}_{2,l,t}} f_W(X_{v,t}) s'(h_u^{(1)}) s(h_{u,j}^{(2)})$$
$$- \sum_{u}^{\mathcal{U}_{2,l,t}} \sum_{v}^{V_{r,t}} Q_v(p_E, p_I) s'(h_u^{(1)}) s(h_{u,j}^{(2)}) \tag{B.8}$$

B.2 The Solution to the Loss $\mathcal{L}_c$

In this section, the solution to the loss function $\mathcal{L}_c$ in Section VI-B of the original paper is described. Specifically, the derivative is solved with respect to W and the scaling parameter λ_2 updated alternately.

The derivative of $\mathcal{L}_c$ with respect to W is as below:

$$\frac{\partial \mathcal{L}_{c,i}}{\partial W_{j,k}^{(1)}} = \sum_{l,t=a_s}^{L,a_e} \sum_{u}^{\mathcal{U}_{2,l,t}} (\lambda_2 \alpha \cdot \sum_{l,p=a_s}^{L,a_e} \sum_{v}^{\mathcal{U}_{2,l,p}} f_W(X_{v,p}) - C(i))$$
$$\cdot s'(h_t^{(1)}) W_j^{(2)} s'(h_{t,j}^{(2)}) X_{u,t,k}^{(n)} \tag{B.9}$$

where $\alpha = (a_e - a_s + 1)$.

$$\frac{\partial \mathcal{L}_{c,i}}{\partial W_j^{(2)}} = \sum_{l,t=a_s}^{L,a_e} \sum_{u}^{\mathcal{U}_{2,l,t}} \lambda_2 \alpha \sum_{l,p=a_s}^{L,a_e} \sum_{v}^{\mathcal{U}_{2,l,p}} f_W(X_{v,p}) s'(h_t^{(1)}) s(h_{t,j}^{(2)}) \\ -C(i) \sum_{l,t=a_s}^{L,a_e} \sum_{u}^{\mathcal{U}_{2,l,t}} s'(h_t^{(1)}) s(h_{t,j}^{(2)}) \tag{B.10}$$

In addition, the analytical solution of the scaling factor λ_2 is as follows:

$$\lambda_2 = \sum_{i}^{T'} M_i \cdot C(i) / \sum_{i}^{T'} M_i^2 \tag{B.11}$$

where $M_i = (a_e - a_s + 1) \sum_{l,t=a_s}^{L,a_e} \sum_{u}^{\mathcal{U}_{2,l,t}} f_W(X_{u,t})$.

B.3 Settings of Comparison Methods

This section introduces the 6 competing methods. Among these, 4 are from social media mining: *Linear Autoregressive Exogenous model (LinARX)* [2], *Logistic Autoregressive Exogenous model (LogARX)* [3], *Simple Linear Regression model (simpleLinReg)* [51], *Multivariable linear regression model (multiLinReg)* [37] and the remaining 2 methods are from computational epidemiology: *SEIR* [76] and *EpiFast* [14]. Their detailed settings are presented in the following.

(1) *Linear Autoregressive Exogenous model (LinARX)* [2,83]: This is a standard ARX model that builds the dependence of future visit percentage on the historical time series of CDC ILI visit percentage data [28] and the volume of influenza tweet data $\mathcal{D}_{(+)}$. The orders of LinARX for Twitter data time series and CDC time series are set as 2 and 3, respectively, based on cross-validation.

(2) *Logistic Autoregressive Exogenous model (LogARX)* [3]: On the basis of LinARX, this method adds a logit function transformation to the historical time series to enforce the boundary 0-1 of the value of ILI visit percentage. The orders of LogARX for the two time series are both set as 2 based on cross-validation.

(3) *Simple Linear Regression model (simpleLinReg)* [51]: This method assumes a linear mapping between the input, the volume of infectious tweets $\mathcal{D}_{(+)}$, and the output, which is the future ILI visit percentage.

(4) *Multi-variable linear regression model (multiLinReg)* [37]: This method treats a combination of keywords $\mathcal{K}$'s volumes as a multivariate input of the simple regression model.

(5) *SEIR* [76]: This model divides the population into four health states, namely susceptible (S), exposed (E), infectious (I), and recovered (R). The epidemic dynamics are modeled by ordinary differential equations. The visit percentage is calculated by multiplying the volume of the state "I" by some ratio, which is optimized by cross-validation.

(6) *EpiFast* [14]: This model follows the definition in Section III, applying two main parameters that must be tuned, p_E and p_I. These are optimized by minimizing the error of the predicted and the actual ILI visit percentages via the Nelder Mead method [18].

Bibliography

[1] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *The Journal of Machine Learning Research*, 10:803–826, 2009.

[2] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. Predicting flu trends using Twitter data. In *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 702–707, 2011.

[3] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. Online social networks flu trend tracker: a novel sensory approach to predict flu trends. In *Biomedical Engineering Systems and Technologies*, pages 353–368. Springer, 2013.

[4] C. C. Aggarwal and K. Subbian. Event detection in social streams. In *ICDM '12*, pages 624–635, 2012.

[5] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Göker, I. Kompatsiaris, and A. Jaimes. Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, 15(6):1268–1282, 2013.

[6] R. M. Anderson and R. M. May. Population biology of infectious diseases: Part i. *Nature*, (280):361–7, 1979.

[7] R. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

[8] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. *Advances in neural information processing systems*, 19:41, 2007.

[9] M. Arias, A. Arratia, and R. Xuriguera. Forecasting with Twitter data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):8, 2013.

[10] A. Bandyopadhyay, K. Ghosh, P. Majumder, and M. Mitra. Query expansion for microblog retrieval. *International Journal of Web Science*, 1(4):368–380, 2012.

[11] C. Barrett, R. Beckman, M. Khan, V. Kumar, M. Marathe, P. Stretz, T. Dutta, and B. Lewis. Generation and analysis of large synthetic social contact networks. In *WSC*, pages 1003–1014, Dec. 2009.

[12] C. Barrett, K. Bisset, S. Eubank, X. Feng, and M. Marathe. Episimdemics: An efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *ICS*, pages 1–12, Nov. 2008.

[13] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[14] R. Beckman, K. R. Bisset, J. Chen, B. Lewis, M. Marathe, and P. Stretz. Isis: A networked-epidemiology based pervasive web app for infectious disease pandemic planning and response. In *KDD*, pages 1847–1856. ACM, 2014.

[15] S. Bennett. Facebook, twitter, instagram, pinterest, vine, snapchat–social media stats 2014 [infographic]. *Retrieved November*, 8:2014, 2014.

[16] C. M. Bishop et al. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[17] K. Bisset, J. Chen, X. Feng, V. S. A. Kumar, and M. Marathe. Epifast: A fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In *ICS*, pages 430–439, 2009.

[18] K. R. Bisset, J. Chen, X. Feng, V. Kumar, and M. V. Marathe. Epifast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In *ICS*, pages 430–439. ACM, 2009.

[19] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML*, pages 113–120. ACM, 2006.

[20] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

[21] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[22] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[23] S. Brennan, A. Sadilek, and H. Kautz. Towards understanding global spread of disease from everyday interpersonal interactions. In *IJCAI*, pages 2783–2789. AAAI Press, 2013.

[24] O. Cappé. Online expectation-maximisation. *Mixtures: Estimation and Applications*, pages 1–53, 2011.

[25] O. Cappé and E. Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, 2009.

[26] C. Carpineto and G. Romano. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1, 2012.

[27] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[28] CDC. Fluview interactive. Accessed May 31, 2015. http://www.cdc.gov/flu/weekly/fluviewinteractive.htm.

[29] D. Chakrabarti and K. Punera. Event summarization using tweets. In *ICWSM '11*, pages 66–73, 2011.

[30] P. Chakraborty, P. Khadivi, B. Lewis, A. Mahendiran, J. Chen, P. Butler, E. O. Nsoesie, S. R. Mekaru, J. S. Brownstein, M. Marathe, et al. Forecasting a moving target: Ensemble models for ILI case count predictions. In *SDM 2014*, pages 262–270, 2014.

[31] C. C. Chen, M. C. Chen, and M.-S. Chen. Liped: HMM-based life profiles for adaptive event detection. In *SIGKDD*, pages 556–561, 2005.

[32] J. Chen, J. Liu, and J. Ye. Learning incoherent sparse and low-rank patterns from multiple tasks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):22, 2012.

[33] L. Chen, K. T. Hossain, P. Butler, N. Ramakrishnan, and B. A. Prakash. Flu gone viral: Syndromic surveillance of flu on Twitter using temporal topic models. In *ICDM*, pages 2783–2789. IEEE, 2014.

[34] N. H. Choi, W. Li, and J. Zhu. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364, 2010.

[35] N. Collier, N. T. Son, and N. M. Nguyen. Omg u got flu? analysis of shared health messages for bio-surveillance. *J. Biomedical Semantics*, 2(S-5):S9, 2011.

[36] M. E. Craft, E. Volz, C. Packer, and L. A. Meyers. Disease transmission in territorial populations: the small-world network of serengeti lions. *Journal of the Royal Society Interface*, 8(59):776–786, 2011.

[37] A. Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122. ACM, 2010.

[38] M. Dredze, M. J. Paul, S. Bergsma, and H. Tran. Carmen: A Twitter geolocation system with applications to public health. In *AAAI Workshop on Expanding the Boundaries of HIAI*, pages 20–24. Citeseer, 2013.

[39] T. Evgeniou and M. Pontil. Regularized multi-task learning. In *KDD 2004*, pages 109–117. ACM, 2004.

[40] S. Gao. A shared random effect parameter approach for longitudinal dementia data with non-ignorable missing data. *Statistics in Medicine*, 23(2):211–219, 2004.

[41] M. S. Gerber. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014.

[42] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. Johns Hopkins University Press, 2012.

[43] P. Gong, C. Zhang, Z. Lu, J. Z. Huang, and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *ICML*, volume 28, page 37. NIH Public Access, 2013.

[44] C. Groendyke, D. Welch, and D. R. Hunter. A network-based analysis of the 1861 hagelloch measles data. *Biometrics*, 68(3):755–765, 2012.

[45] S. E. Hardy, H. Allore, and S. A. Studenski. Missing data: A special challenge in aging research. *Journal of the American Geriatrics Society*, 57(4):722–729, 2009.

[46] A. Haris, D. Witten, and N. Simon. Convex modeling of interactions with strong heredity. *arXiv preprint arXiv:1410.3517*, 2014.

[47] F. E. Harrell. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer Science & Business Media, 2013.

[48] T. H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *TKDE '03*, 15(4):784–796, 2003.

[49] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence. Improving traffic prediction with tweet semantics. In *IJCAI*, pages 1387–1393, 2013.

[50] J. M. Hernandez-lobato, N. Houlsby, and Z. Ghahramani. Probabilistic matrix factorization with non-random missing data. In *ICDM 2014*, pages 1512–1520, 2014.

[51] H. Hirose and L. Wang. Prediction of infectious disease spread using Twitter: A case of influenza. In *PAAP*, pages 100–105. IEEE, 2012.

[52] L. Hong, B. Dom, S. Gurumurthy, and K. Tsioutsiouliklis. A time-dependent topic model for multiple text streams. In *KDD*, pages 832–840. ACM, 2011.

[53] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *ICML 2009*, pages 457–464. ACM, 2009.

[54] F. Jin, R. P. Khandpur, N. Self, E. Dougherty, S. Guo, F. Chen, B. A. Prakash, and N. Ramakrishnan. Modeling mass protest adoption in social network communities using geometric brownian motion. In *KDD 2014*, pages 1660–1669. ACM, 2014.

[55] F. Jin, W. Wang, L. Zhao, E. Dougherty, Y. Cao, C.-T. Lu, and N. Ramakrishnan. Misinformation propagation in the age of twitter. *Computer*, (12):90–94, 2014.

[56] V. R. Joseph. A Bayesian approach to the design and analysis of fractionated experiments. *Technometrics*, 48(2):219–229, 2006.

[57] N. Kallus. Predicting crowd behavior with big public data. In *WWW 14 Companion*, pages 625–630. IW3C2, 2014.

[58] M. Krieck, J. Dreesman, L. Otrusina, and K. Denecke. A new age of public health: Identifying disease outbreaks by analyzing tweets. In *WebSci*, 2011.

[59] M. Kulldorff. A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496, 1997.

[60] G. Kumaran and J. Allan. Text classification and named entities for new event detection. In *SIGIR '04*, pages 297–304, 2004.

[61] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW*, pages 591–600, 2010.

[62] A. Lamb, M. J. Paul, and M. Dredze. Separating fact from fear: Tracking flu infections on Twitter. In *HLT-NAACL*, pages 789–795, 2013.

[63] T. Lappas, M. R. Vieira, D. Gunopulos, and V. J. Tsotras. On the spatiotemporal burstiness of terms. *VLDB '12*, 5(9):836–847, 2012.

[64] R. Li, K. H. Lei, R. Khadiwala, and K.-C. Chang. TEDAS: a Twitter-based event detection and analysis system. In *ICDE*, pages 1273–1276. IEEE, 2012.

[65] R. Li, S. Wang, and K. C.-C. Chang. Towards social data platform: automatic topic-focused monitor for Twitter stream. *VLDB*, 6(14):1966–1977, 2013.

[66] M. Lim and T. Hastie. Learning interactions through hierarchical group-lasso regularization. *arXiv preprint arXiv:1308.2719*, 2013.

[67] C. X. Lin, Q. Mei, J. Han, Y. Jiang, and M. Danilevsky. The joint inference of topic diffusion and evolution in social communities. In *ICDM*, pages 378–387. IEEE, 2011.

[68] J. Lin, R. Snow, and W. Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *KDD*, pages 422–429. ACM, 2011.

[69] L. Manelzis and S. Peleg. War journalism as media manipulation: Seesawing between the Second Lebanon war and the Iranian nuclear threat. *Peace and Policy*, 13:62–73, 2008.

[70] M. Marchetti-Bowick and N. Chambers. Learning for microblogs with distant supervision: Political forecasting with Twitter. In *ECACL*, pages 603–612, 2012.

[71] K. Massoudi, M. Tsagkias, M. de Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. In *Advances in Information Retrieval*, pages 362–367. Springer, 2011.

[72] S. Mehrotra. On the implementation of a primal-dual interior point method. *SIAM Journal on Optimization*, 2(4):575–601, 1992.

[73] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD*, pages 198–207. ACM, 2005.

[74] MITRE. http://www.mitre.org/.

[75] S. Muff, F. Rao, and A. Caflisch. Local modularity measure for network clusterizations. *Physical Review E*, 72(5):056107, 2005.

[76] J. D. Murray. Mathematical biology I: An introduction. *Interdisciplinary Applied Mathematics*, 17, 2002.

[77] D. B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):337–360, 2012.

[78] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129, 2010.

[79] J. O. Ogutu, T. Schulz-Streeck, and H.-P. Piepho. Genomic selection using regularized linear regression models: ridge regression, Lasso, elastic net and their extensions. In *BMC proceedings*, volume 6, page S10. BioMed Central Ltd, 2012.

[80] PAHO. Paho interactive. Accessed May 31, 2015. www.paho.org/hq/.

[81] M. J. Paul and M. Dredze. A model for mining public health topics from Twitter. *Health*, 11:16–6, 2012.

[82] M. J. Paul and M. Dredze. Discovering health topics in social media using topic models. *PloS one*, 9(8):e103408, 2014.

[83] M. J. Paul, M. Dredze, and D. Broniatowski. Twitter improves influenza forecasting. *PLoS currents*, 6, 2014.

[84] B. Pavlyshenko. Forecasting of events by tweet data mining. *arXiv preprint arXiv:1310.3499*, 2013.

[85] A. M. Presanis, D. De Angelis, A. Hagy, C. Reed, S. Riley, B. S. Cooper, L. Finelli, P. Biedrzycki, M. Lipsitch, et al. The severity of pandemic H1N1 influenza in the united states, from april to july 2009: a bayesian analysis. *PLoS medicine*, 6(12):e1000207, 2009.

[86] 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 *SIGKDD*, pages 1799–1808. ACM, 2014.

[87] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. In *ICWSM*, pages 297–304, 2011.

[88] A. Ritter, O. Etzioni, S. Clark, et al. Open domain event extraction from Twitter. In *KDD '12*, pages 1104–1112, 2012.

[89] J. Ritterman, M. Osborne, and E. Klein. Using prediction markets and Twitter to predict a swine flu pandemic. In *1st international workshop on mining social media*, volume 9, 2009.

[90] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW '10*, pages 851–860, 2010.

[91] E. Schubert, M. Weiler, and H.-P. Kriegel. Signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds. In *KDD 2014*, pages 871–880. ACM, 2014.

[92] A. Signorini, A. M. Segre, and P. M. Polgreen. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS one*, 6(5):e19467, 2011.

[93] S. Thrun and J. O'Sullivan. Clustering learning tasks and the selective cross-task transfer of knowledge. *Learning to learn*, pages 181–209, 1998.

[94] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[95] Z. Tufekci and C. Wilson. Social media and the decision to participate in political protest: Observations from Tahrir Square. *Journal of Communication*, 62(2):363–379, 2012.

[96] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[97] E. Vynnycky and R. G. White. *An Introduction to Infectious Disease Modelling*. Oxford University Press, 2010.

[98] X. Wang, D. E. Brown, and M. S. Gerber. Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information. In *ISI*, pages 36–41, 2012.

[99] X. Wang, M. S. Gerber, and D. E. Brown. Automatic crime prediction using events extracted from Twitter posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 231–238. Springer, 2012.

[100] J. Weng and B.-S. Lee. Event detection in Twitter. In *ICWSM '11*, pages 401–408, 2011.

[101] WHO. Ebola data and statistics. Accessed May 29, 2015. http://apps.who.int/gho/data/view.ebola-sitrep.ebola-summary-latest.

[102] WHO. Influenza (season) fact sheet. Accessed May 15, 2015. http://www.who.int/mediacentre/factsheets/fs211/en/.

[103] C. Wilson and A. Dunn. Digital media in the Egyptian revolution: Descriptive analysis from the Tahrir data sets. *International Journal of Communication*, 5:1248–1272, 2011.

[104] S. Xiang, T. Yang, and J. Ye. Simultaneous feature and feature group selection through hard thresholding. In *KDD 2014*, pages 532–541. ACM, 2014.

[105] X. Yang, A. Ghoting, Y. Ruan, and S. Parthasarathy. A framework for summarizing and analyzing Twitter feeds. In *KDD*, pages 370–378. ACM, 2012.

[106] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical topic discovery and comparison. In *WWW '11*, pages 247–256, 2011.

[107] H. Yu, C. Ho, Y. Juan, and C. Lin. Libshorttext: A library for short-text classification and analysis. Technical report, Technical Report. http://www. csie. ntu. edu. tw/˜ cjlin/ papers/libshorttext. pdf, 2013.

[108] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, and J. Ye. Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. In *KDD 2012*, pages 1149–1157. ACM, 2012.

[109] L. Zhao, F. Chen, J. Dai, T. Hua, C.-T. Lu, and N. Ramakrishnan. Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling. *PloS one*, 9(10):e110206, 2014.

[110] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan. Spatiotemporal event forecasting in social media. In *SDM 15*, pages 963–971. SIAM, 2015.

[111] L. Zhao, J. Chen, F. Chen, W. Wang, C.-T. Lu, and N. Ramakrishnan. Simnest: Social media nested epidemic simulation via online semi-supervised deep learning. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 639–648. IEEE, 2015.

[112] 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*, pages 1503–1512. ACM, 2015.

[113] J. Zhou, J. Chen, and J. Ye. *MALSAR: Multi-tAsk Learning via StructurAl Regularization*. Arizona State University, 2011.