INTERPRETABLE SPATIO-TEMPORAL NETWORKS FOR MODELING AND FORECASTING SOCIETAL EVENTS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ALİ MERT ERTUĞRUL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

NOVEMBER 2019

Approval of the thesis:

INTERPRETABLE SPATIO-TEMPORAL NETWORKS FOR MODELING AND FORECASTING SOCIETAL EVENTS

submitted by **ALİ MERT ERTUĞRUL** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Information Systems Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics**

Prof. Dr. Sevgi Özkan Yıldırım
Head of Department, **Information Systems**

Assoc. Prof. Dr. Tuğba Taşkaya Temizel
Supervisor, **Information Systems Dept., METU**

Assoc. Prof. Dr. Yu-Ru Lin
Co-supervisor, **School of Comp. and Inf., University of Pittsburgh**

Examining Committee Members:

Assoc. Prof. Dr. Banu Günel Kılıç
Information Systems Dept., METU

Assoc. Prof. Dr. Tuğba Taşkaya Temizel
Information Systems Dept., METU

Assist. Prof. Dr. Ramazan Gökberk Cinbiş
Computer Engineering Dept., METU

Assist. Prof. Dr. Hamdi Dibeklioğlu
Computer Engineering Dept., Bilkent University

Assist. Prof. Dr. Hacer Yalım Keleş
Computer Engineering Dept., Ankara University

Date: 20.11.2019

Name, Surname: ALİ MERT ERTUĞRUL

Signature :

ABSTRACT

INTERPRETABLE SPATIO-TEMPORAL NETWORKS FOR MODELING AND FORECASTING SOCIETAL EVENTS

ERTUĞRUL, ALİ MERT
Ph.D., Department of Information Systems
Supervisor: Assoc. Prof. Dr. Tuğba Taşkaya Temizel
Co-Supervisor: Assoc. Prof. Dr. Yu-Ru Lin

November 2019, 105 pages

The relationships between individual activities and societal events (e.g. migrations, social movements) are complex due to the various social, temporal and spatial factors. Understanding such relationships in the context of various societal events such as street protests and opioid crisis, and forecasting these events are important as they have great impacts on public policies and supporting decision making of authorities. In this thesis, novel, spatio-temporal, deep neural networks are proposed (i) to forecast societal events and (ii) to help examine the relationships between societal events and their social and geographical contexts. The proposed models are designed to model the complex interactions between local (observed from within a location) and global (observed from all locations) activities by incorporating a new design of attentional networks. They are capable of forecasting the occurrence of future societal events and allow for interpreting what features, from which places, and how they contribute to event forecasting. Within the scope of this thesis, extensive experiments are conducted to evaluate the proposed networks on two different types of population-level societal events, namely social movements and opioid overdoses, with multiple datasets. The results indicate that the proposed models achieve superior forecasting performance than the compared methods. Also, they provide meaningful interpretations in terms of (i) what local and global activity features are more predictive, (ii) what locations have more salient contributions, and (iii) how these locations contribute to forecasting the subsequent events.

Keywords: Spatio-Temporal Learning, Deep Learning, Interpretable Learning, Societal Event Forecasting, Attentional Networks

ÖZ

TOPLUMSAL OLAYLARIN MODELLENMESİ VE TAHMİNİ İÇİN YORUMLANABİLİR UZAY-ZAMANSAL AĞLAR

ERTUĞRUL, ALİ MERT
Doktora, Bilişim Sistemleri Bölümü
Tez Yöneticisi: Doç. Dr. Tuğba Taşkaya Temizel
Ortak Tez Yöneticisi: Doç. Dr. Yu-Ru Lin

Kasım 2019 , 105 sayfa

Bireysel etkinlikler ve toplumsal olaylar (örneğin göçler, sosyal hareketler gibi) arasındaki ilişkiler, çeşitli sosyal, zamansal ve mekansal faktörler nedeniyle karmaşıktır. Bu tür ilişkileri, sokak protestoları ve opioid krizleri gibi çeşitli toplumsal olaylar bağlamında anlamak ve bu olayları tahmin etmek, kamu politikaları ve yetkililerin karar vermelerini desteklemede önemlidir. Bu tezde, (i) toplumsal olayları öngörmek ve (ii) toplumsal olaylar ile onların sosyal ve coğrafi bağlamları arasındaki ilişkileri incelemeye yardımcı olmak için yeni, uzay-zamansal, derin öğrenme ağları önerilmektedir. Önerilen modeller, yeni bir dikkat çeken ağ tasarımı ile yerel (bir konumdan gözlenen) ve küresel (tüm konumlardan gözlenen) faaliyetler arasındaki karmaşık etkileşimleri modellemektedir. Bu modeller, gelecekteki toplumsal olayların oluşumunu öngörebilir ve hangi özniteliklerin, hangi yerlerden ve olay tahminine nasıl katkıda bulunduklarını yorumlamaya izin verirler. Bu tez kapsamında, önerilen ağları değerlendirmek için, toplumsal hareketler ve opioid doz aşımı olmak üzere iki farklı popülasyon düzeyinde toplumsal olayda, çoklu veri setleri üzerinde kapsamlı deneyler yapılmıştır. Sonuçlar, önerilen modellerin karşılaştırılan yöntemlerden daha iyi tahmin performansı sağladığını göstermektedir. Ayrıca, önerilen modeller (i) hangi yerel ve küresel faaliyet özniteliklerin daha öngörücü olduğu, (ii) hangi konumların daha belirgin katkıları olduğu ve (iii) bu konumların daha sonraki olayları tahmin etmeye nasıl katkıda bulunduğu hakkında anlamlı yorumlar sağlamaktadır.

Anahtar Kelimeler: Mekansal-Zamansal Öğrenme, Derin Öğrenme, Yorumlanabilir Öğrenme, Toplumsal Olay Tahmini, Dikkat Ağları

To my family

ACKNOWLEDGMENTS

This may be the hardest part of this thesis to write. Now, it is time to thank every single person touching my life to complete this long journey!

I would like to first express my sincere gratitude to my supervisor Assoc. Prof. Dr. Tuğba Taşkaya Temizel. Not only being guiding and supportive but also encouraging me when I fell into desperation, she always instilled confidence in me throughout the thesis period. Her enlightening ideas helped me a lot during every step of this research. With my deepest gratitude, I would like to thank my co-supervisor Assoc. Prof. Dr. Yu-Ru Lin for every single moment of this thesis period starting from her accepting me to PICSO LAB. I have always felt privileged to be able to work with her. With her stance in academia, enthusiasm to learn, extraordinary vision, endless support and bright ideas that shaped my research, she has been more than a supervisor, a role model indeed. It was an honour for me to be supervised by these two great people throughout my thesis period.

I am also thankful to the thesis supervising committee members Assist. Prof. Dr. Gökberk Cinbiş and Assoc. Prof. Dr. Banu Günel Kılıç for their invaluable comments and assistance for helping me accomplish this dissertation. I would also like to thank the thesis examining committee members Assist. Prof. Dr. Hamdi Dibeklioğlu and Assist. Prof. Dr. Hacer Yalım Keleş for their valuable comments and feedback to improve this dissertation.

I would like to thank the special members of the PICSO LAB; Muheng Yan, Yongsu Ahn, Xian Teng, Xidao Wen, Wen-Ting Chung, Ang Li who kindly and warmly welcomed me, and contributed to this thesis a lot with their invaluable discussions and precious suggestions. It was a great pleasure for me to work with you guys!

I would also like to thank my dear friends; Serhat Peker, Emre Sezgin, Şeyma Çavdar, Mehmet Ali Akyol and Sibel Gülnar for their encouragements, continuous support as well as the joyful moments and sweet coffee breaks at the institute. I am also thankful to my amazing friends from Pitt; Nuray Baltacı Akhüseyinoğlu, Kamil Akhüseyinoğlu, Jordan Barría Pineda, Marcin Koźniewski, Jidapa Kraisangka and Fan Yang for sharing beautiful memories and our laughter-filled conversations.

I would like to express my profound appreciation to my wonderful family. Their endless love, compassion, peace, never ending support and encouragement enabled me to take firm steps to where I am today. They never stopped believing in me and I know they will be always there whenever I need. I am truly grateful to my mother Fulya, my father Dursun and my lovely sister Naz. I am also thankful to my dearest parents-in-law Deniz and Mustafa, and my brother-in-law Mert for always caring me, motivating me and making me feel loved.

Last but surely not the least, my deepest sense of appreciation goes to my beloved and beautiful wife Itır. Words cannot describe how thankful and blessed I am for having her with me in this life, in this adventure. Her everlasting support, endless love, adoring glances and heartwarming smile turned the miserable and bitter moments into warm and sweet memories not

only during my Ph.D. journey but also throughout my life. We cried, worried about the future, laughed like a drain, discovered the beauties of the world, and grew up together. This chapter of our lives has already gone. Now, we are opening a newer one which may be shot in another city, even in another continent of the world. To our new adventures, new joyful memories! Cheers my darling!

I acknowledge the support from PICSO LAB and the University of Pittsburgh through fundings including NSF #1634944, #1637067, #1739413, DARPA UGB, AFOSR, and the University of Pittsburgh ULS Open Access Author Fee Fund.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

OUD	Opioid Use Disorders
MLP	Multi-layer Perceptrons
NN	Neural Network
DNN	Deep Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
CNN	Convolutional Neural Network
SAE	Stacked Auto Encoders
NMT	Neural Machine Translation
NLP	Natural Language Processing
ADMM	Alternating Direction Method by Multiplier
BPTT	Back Propagation Through Time
HMM	Hidden Markov Model
FPMC	Factorizing Personalized Markov Chain
TF	Tensor Factorization
LDA	Latent Dirichlet Allocation
SGD	Stochastic Gradient Descent
BLM	Black Lives Matter
LR	Logistic Regression
SVM	Support Vector Machine
RMTFL	Regularized Multi-Task Feature Learning
CMTFL-1	Constrained Multi-Task Feature Learning I

CMTFL-2	Constrained Multi-Task Feature Learning II
GL	Group Lasso
AUC	Area Under Curve
EMS	Emergency Medical Service
SNA	Statistical Neighborhood Approximation
SVR	Support Vector Regression
VAR	Vector Autoregression
ARIMA	Autoregressive Integrated Moving Average

CHAPTER 1

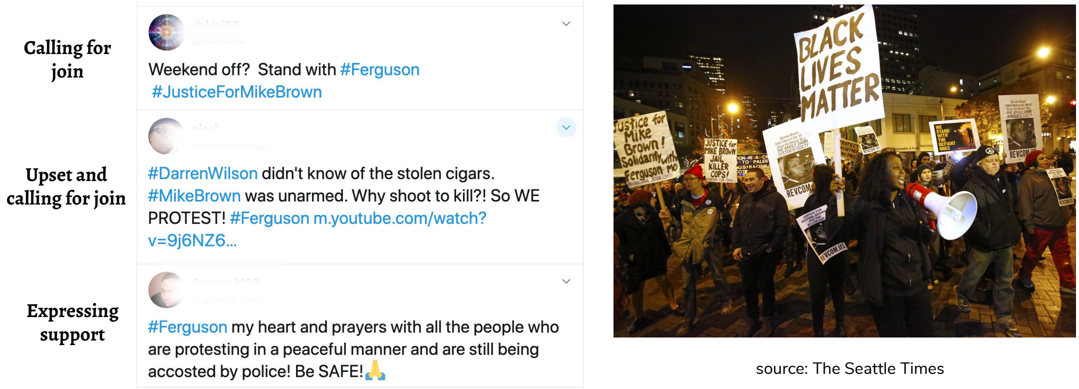
INTRODUCTION

1.1 Motivation

A societal event can be described as an event, which is of high importance to the society such as migration, social movement, and experience of a specific health-related event. Societal events are associated with the temporal and spatial sequences of individual actions, which follow patterns in particular types of environments. Depending on various social, temporal and spatial factors, the relationships between such individual activities and societal events might be complex. Understanding these relationships in the context of various societal events such as street protests and opioid crisis, and predicting future events are important as they have great impacts on public policies and supporting decision making of authorities [1, 2].

Figure 1.1 indicates an example of online activities of the individuals on social media, in particular Twitter, where they talk about after shooting of unarmed Michael Brown by a police officer in Ferguson, Missouri. Following the shooting event, Ferguson unrests occurred, which are the symbolic protests against systemic racism to Black people in the U.S. Later on, the Black Lives Matter (BLM) movement was nationally recognized [3]. As shown in the example, people give voice to their grievances and concerns about happenings, call for join to future protests, express their support to the protesters through social media. In this case, online activities of the individuals are connected and interrelated to the offline protests. Furthermore, Figure 1.2a reveals a tweet posted by a police and crime commissioner in the U.K. stating the connection between criminal activities, particularly theft, burglary and robbery, and opioid use such as heroin and cocaine. Similarly, Figure 1.2b indicates the relationship between involvement in criminal justice system and intensity of opioid use based on a recent published study [4]. It is stated that involvement in the criminal justice system increases with intensity of opioid use, and opioid use is significantly associated with involvement in the criminal justice system. These examples highlight the potential links between the individual criminal activities and the opioid crisis. Given the potential links and the interrelations between individual activities and societal events, how to uncover these relationships and predict their future behaviors have drawn significant research attention across multiple disciplines such as information science [1], public health [5] and political science [6]. However, this task is challenging due to complex nature of these relationships.

The drastic increase in availability of digital traces of activities from various data sources (e.g. social media, city sensors, computer-aided dispatch databases), and recent improvements in predictive analytics offer a powerful way to learn the relationships between activities and


Calling for join
Upset and calling for join
Expressing support
BLACK LIVES MATTER
Justice for Mike Brown! Solidarity with Ferguson MO

Weekend off? Stand with #Ferguson #JusticeForMikeBrown

#DarrenWilson didn't know of the stolen cigars. #MikeBrown was unarmed. Why shoot to kill?! So WE PROTEST! #Ferguson m.youtube.com/watch?v=9j6NZ6...

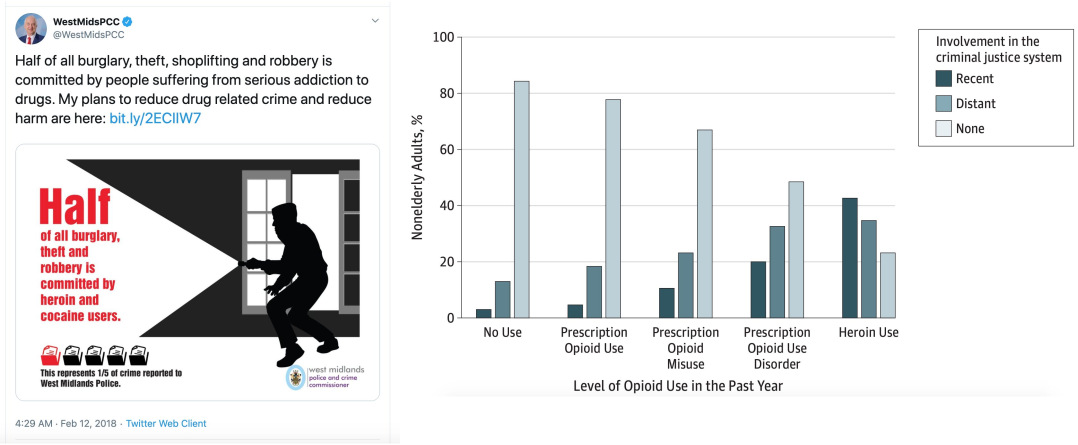
#Ferguson my heart and prayers with all the people who are protesting in a peaceful manner and are still being accosted by police! Be SAFE!🙏

source: The Seattle Times

Figure 1.1: Sample tweets posted after shooting of Michael Brown in Ferguson, Missouri (left) in 2014 and a photo from a protest following the shooting event (right).

societal events. However, they often provide "black box" solutions and give no or little insight into interpreting learned relationships [7, 8, 9, 10]. A predictive model with interpretation capability in terms of these relationships can give insights and enable related stakeholders to forecast future events and be prepared for them. For instance, stakeholders such as police departments or event planners can plan and allocate resources to encourage peaceful and orderly crowd gatherings, marches or demonstrations. Similarly, forecasting opioid overdoses and uncovering the underlying reasons of opioid use enable governments (e.g. public health) to improve the overdose surveillance and to identify the areas in need of prevention effort considering the spatial differences.

Spatio-temporal modeling and forecasting of societal events from digital traces of the activities are important problems but encounters critical challenges. (1) The massive amount of dig-


WestMidsPCC
@WestMidsPCC

Half of all burglary, theft, shoplifting and robbery is committed by people suffering from serious addiction to drugs. My plans to reduce drug related crime and reduce harm are here: bit.ly/2ECIlW7

4:29 AM · Feb 12, 2018 · Twitter Web Client

(a)

(b)

Figure 1.2: Examples showing the connection between criminal activities and opioid use. (a) Tweet stating the relationship between specific types of criminal activities and dependents of opioid use by a police and crime commissioner in the U.K. (b) Relationship between involvement in criminal activities and opioid use, published by [4].

ital activity traces surrounding various activities calls for the need of an automatic approach to learn patterns that reflect differences in their social, spatial and temporal contexts. How to automatically learn the heterogeneous contexts in a unified framework is an open question. (2) Such automatic approach needs to enable an interpretation about how events unfold over time and across space which shows the learned relationships between activities and the events. In this context, the main focus of this thesis is to seek to address the aforementioned problems by proposing novel, interpretable predictive deep networks to model and forecast the societal events. To evaluate the proposed networks, extensive experiments are conducted on two different types of population-level societal events, namely social movements and opioid overdoses, with multiple datasets.

1.2 Research Questions

This thesis aims to answer two main research questions as follows:

- **RQ 1)** How can a method be developed that learns the relationships between the activities and societal events, and forecasts the future events from the learned patterns that reflect differences in social, spatial and temporal contexts?
- **RQ 2)** To what extent does the automatic method allow for interpretation about how events unfold over time and across space?
  - **RQ 2.1)** What social and activity features are associated with the subsequent events?
  - **RQ 2.2)** To what extent are the local activities (observed from within a region) and the global activities (observed outside of a region) predictive to the subsequent events?
  - **RQ 2.3)** What locations' activities would have a more far-reaching predictive power, in terms of signaling subsequent events in other locations?

Within the scope of this thesis, two main studies are conducted to answer the research questions above. Each study seeks to answer all research questions. Each study follows a similar pipeline which includes a novel interpretable predictive model that tackles the drawbacks of existing predictive models, data collection, feature extraction and evaluation of the proposed model on a specific societal event domain.

First, the existing spatio-temporal predictive models are reviewed and their shortcomings are identified. Accordingly, a novel interpretable spatio-temporal predictive model is proposed that learns the relationships between individual activities and societal events, and forecasts the future events. The proposed model is evaluated on the domain of social movements with three different cases, where the purpose is to forecast future offline protests from spatio-temporal online activities. The informative local and global activity features, contributions of local and activity features as well as the most contributing locations (hubs) to forecasting are interpreted as a result of the structure of the proposed model.

Second, a novel, community-attentive, spatio-temporal and interpretable predictive model is proposed to model the relationships between activities and societal events, and forecast the future societal events in an interpretable way. The proposed model is evaluated on the domain of opioid crisis, where the purpose is to forecast future opioid overdoses from spatio-temporal crime activities. The important local-level and global-level activity features, learned communities, community memberships and community contributions to forecasting future local events are analyzed and interpreted.

1.3 Main Contributions

The contributions of this thesis are summarized as follows:

- **A unified, spatio-temporal predictive network:** A novel spatio-temporal predictive deep learning model, called *ActAttn*, is proposed, which (i) automatically learns the relationships between the spatio-temporal activity traces and societal events, and (ii) forecasts the future events. To the best of our knowledge, this is the first model that differentiates the local and global contributions in the spatio-temporal societal event forecasting domain.

- **A community-attentive spatio-temporal predictive network:** A novel multi-head attention based deep predictive model, called *CASTNet*, is proposed, which learns different representation subspaces of global dynamics (i.e. communities). *CASTNet* allows the prediction for a given location to be individually optimized by the features contributed by a mixture of communities to effectively forecast the societal events.

- **Interpretability in hierarchical attention:** Hierarchical attentional networks are incorporated into the proposed models. Through attentional networks, *ActAttn* (i) differentiates the local and global feature contributions and (ii) identifies the hub regions that have a more salient contribution in predicting future societal events. Therefore, it allows for interpreting the importance of activities in different regions (intra- vs. inter-region contribution, and hubs) in forecasting future events. *CASTNet* allows for interpreting (i) learned communities, (ii) community memberships (which locations form the communities), (iii) community contributions for forecasting local incidents, and (iv) informative time steps in both local-level and global-level.

- **Interpretability in activity features:** The proposed models incorporate Group Lasso (GL) regularization to select informative set of features which succinctly capture what activity types at both local- and global-level are more associated with the future societal events.

- **Investigation of opioid overdose forecasting capability from crime dynamics:** To the best of our knowledge, for the first time, this thesis proposes to forecast future opioid overdoses from the spatio-temporal crime dynamics and location-specific features.

- **Extensive experiments:** Extensive experiments and in-depth analyses are conducted to evaluate each proposed model. *ActAttn* is evaluated on three real-world social movement datasets. Moreover, the applicability of *CASTNet* is investigated by conducting experiments on two real-world opioid overdose datasets.

1.4 Organization of the Thesis

The outline of this thesis is as follows:

Chapter 2 provides background information about societal events with a specific focus on social movements and opioid overdose. It continues with comprehensive literature review and state-of-the-art knowledge about the temporal/spatio-temporal prediction/forecasting approaches. Finally, it ends with presenting basic machine learning concepts used in this thesis.

Chapter 3 proposes a novel interpretable spatio-temporal predictive model, called ActAttn, for modeling and forecasting future societal events. It also presents evaluation and in-depth analysis of the proposed model on the domain of social movements.

Chapter 4 proposes a novel community-attentive spatio-temporal predictive model, called CASTNet, for modeling and forecasting future societal events. This chapter also provides evaluation and in-depth analysis of the proposed model on the domain of opioid overdoses.

Chapter 5 summarizes the proposed models and findings, and concludes the thesis with the limitations of these models as well as the possible future directions.

CHAPTER 2

LITERATURE REVIEW AND BACKGROUND

This chapter begins with providing background information about societal events and possible links between them and human activities based on social theories. Then, it continues with a survey of the relevant and state-of-the-art studies for temporal/spatio-temporal prediction/-forecasting by categorizing them into two main groups, deep neural network (DNN) based approaches and other approaches. Finally, this chapter ends with presenting basic machine learning concepts that are used in the proposed methods through this thesis.

2.1 Societal Events

This section provides background information about societal events, in particular two main significant population-level societal events, namely social movements and opioid overdoses. It also presents possible links between the human activities and the corresponding societal events based on relevant social theories. Particularly, it first discourses the relationships between online human-activities and offline social movements. Then, it unveils the relationships between a social phenomena "crime" and drug use.

2.1.1 Social Movements

Social movements are one of the most complex societal events. They reflect how collectivities articulate and press a collectivity's interests to make significant changes in public policies and political decisions. Every day, news about social movement activity relevant to a variety of contested issues is being updated, on topics ranging from civil rights, to human rights, to gender equality, to gun control and others. Throughout human history, protests have been primary means of engaging in social movements, in which collectivities usually give voice to their grievances and concerns about the rights and well-being of themselves and others [11]. In recent decades, the diffusion of new information and communication technologies—social media in particular—has reshaped the political activism of our time. From the Arab Spring, to the Occupy Wall Street movement, to the recent March for Our Lives gun violence protests, social media has been central in providing mobilizing information, coordinating demonstrations, and creating opportunities for people to exchange opinions [12, 13].

Literature in social movements and social psychology has proposed theories and offered insights into why people protest [14, 15, 16]. For instance, one fundamental factor of a given movement is its “connectedness”, both in terms of how events temporally and spatially connect with other events of a similar kind, and in terms of how they are embedded in an environment where people share similar sociocultural context. In other words, social movements are not merely instances of independent collective actions or protest events, but need to be investigated within their social, temporal and geographical contexts [11]. Furthermore, Van Stekelenburgh and Klandermants [14, 15] proposed a motivational framework that incorporates several sociopsychological factors that have been theorized and studied as critical to protests: (1) Identity: individuals’ identification with certain groups/communities brings about a shared sense of future destiny and social responsibility; (2) grievance: a felt sense of illegitimate inequality; (3) emotion: emotions such as anger, guilt, fear, shame, and despair that amplify the felt grievance to be stronger and accelerate people to act more promptly; (4) social embeddedness: the social contexts one is exposed to and social networks one is embedded in, and (5) efficacy: how one perceives that protests could make a difference. Such social embeddedness transforms individual grievance and emotion into their collective forms and may further facilitate the social actions of protests. Therefore, based on the aforementioned social theories, spatio-temporal evolution of social media users’ individual posting behaviors in terms of the given sociopsychological factors can be utilized for the prediction of possible future offline social movements.

Studying social movements through digital platforms has drawn a significant research interest. Among them, Conover et al. [17] examined the temporal evolution of digital communication activity related to the Occupy Wall Street movement using Twitter-centric features. Chung et al. [18] studied online social media discussions during the 2014 Ferguson protests, and employed a thematic analysis to differentiate tweets that engaged critical sense making from those solely focused on the event itself. De Choudhury et al. [19] studied the temporal characteristic of social media participation and its relationships to offline protests related to Black Lives Matter (BLM) movement. However, studies often analyze single events or movements via a case-study approach [17, 20, 21, 22, 18], or consider a large number of movement-related events independently of their relationships in time and space [19, 23].

2.1.2 Opioid Overdose

Opioid use disorders (OUD) and overdose rates in the United States have increased at an alarming rate since the past decade [24]. Overdose deaths involving prescription opioids have been continuously rising since the 1990s; heroin overdose deaths have sharply increased since 2010 [25, 26]. The age-adjusted rate for drug poisoning deaths involving heroin nearly quadrupled between 2000 and 2013 [27], and deaths from drug overdose are now the top cause of injury-related death in the United States [28]. The rate of growth of OUD and overdose, combined with the number of impacted individuals in the United States, has led many to classify this as an “opioid epidemic” [29]. Enhanced understanding of the dynamics of the overdose epidemic may help policy-makers to develop more effective epidemic prevention mechanisms and control strategies [30].

The opioid epidemic is a complex social phenomenon involving and interacting with various social, spatial and temporal factors [5]. Highlighting the links between opioid use and various phenomena and contextual factors has drawn significant research attention including crime and economic stressors. Hammersley et al. [31] stated that opportunities for drug use increase with involvement in criminal behavior. The people dependent on opiates are disproportionately involved in criminal activities [32] especially for the crimes committed for financial gain [33]. Seddon et al. [34] revealed that crime and drug use share common set of causes and they co-occur together. Beside, crime occurrences also have non-trivial spatio-temporal characteristics –for instance, routine activity theory suggested that crimes may exhibit spatio-temporal lags as the likely offenders of one place may reach suitable targets in other places. Therefore, how to unveil the complicated relationship between opioid use and crime incidents is challenging. Most of the existing works studying the relationship between opioid use and social phenomena have employed basic statistical analysis, and focused on current situation and trends rather than predicting/forecasting overdose. Moreover, these studies overlooked the interactions among spatio-temporal dynamics of the locations. Among the studies predicting/forecasting opioid overdose, regression-based approaches have been applied in individual-level [35] and state-level [31].

Detailed assessments of opioid use disorders and overdose growth require systematically collected well-resolved spatio-temporal data [36, 37]. However, the amount of systematically monitored data either at a regional or local level in the U.S. is very limited. Therefore, collecting spatio-temporal well-resolved data is crucial for the assessment of opioid overdose. In addition, there is no common reporting mechanism for incidents. For instance, the incident categories and the organization of categories vary significantly across the databases. On the other hand, crime data is meticulously collected, organized and stored, at a finer-grained level. Given the plausible relationship between the crime dynamics and opioid use as well as the availability of real-time crime data for various locations, modeling spatio-temporal crime dynamics can be a good approach to forecast future opioid overdose.

2.2 Temporal/Spatio-Temporal Prediction/Forecasting Approaches

In the literature, there have been a number of methods/studies that utilize the temporal or spatio-temporal dependencies in modeling and prediction/forecasting. This chapter reviews the relevant works, which are applied or proposed not only in societal event domain, but also in the other domains including time series, meteorology, traffic, etc. Within the scope of this thesis, these studies are divided into two main categories, which are DNN-based approaches and the other approaches. The DNN-based approaches include the studies incorporating neural network-based architectures such as Multilayer Perceptrons (MLP), Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). On the other hand, the other approaches correspond to point processes, Markov-based methods, traditional statistical classification and regression-based models and more advanced methods other than deep learning methods.

2.2.1 DNN-based Approaches for Temporal/Spatio-Temporal Prediction/Forecasting

DNNs have become increasingly popular for temporal and spatio-temporal prediction/forecasting with their state-of-the-art performances on various real-world problems such as neural machine translation (NMT) [38], emotion detection [39] in natural language processing (NLP); and facial action unit detection [40], video classification [41] in computer vision domains. For spatio-temporal modeling and prediction/forecasting problems, CNNs are generally used for modeling temporal (convolution through time) and/or spatial (for grid-like data) relationships while RNNs are employed as the main building blocks for modeling temporal dependencies. The studies reviewed in this section are classified under two categories namely event prediction/forecasting and time-series prediction/forecasting studies.

2.2.1.1 Event Prediction/Forecasting

Among the studies that learn temporal/spatio-temporal dynamics for predicting/forecasting, Hu et al. [42] proposed an hierarchical LSTM (Long Short-Term Memory) encoder-decoder architecture to predict the next sub-event in an event based on the assumption that each event composes sub-events. Given the textual descriptions of previous sub-events, the model predicts and generates the next sub-event description. The first level of the encoder is responsible for sub-event level encoding and the second level encodes the temporal dependency between the sub-events. The decoder part, predicts the next sub-event and generates its description using the encoding of sub-events. Similarly, Granroth-Wilding et al. [43] suggested a compositional neural network model for predicting whether two events are expected to appear in the same chain. The model first embeds the event descriptions in the word-level. Then, it suggests a function that composes the word embeddings into an event representation. In the final step, the model predicts whether given two events are expected to appear in the same chain. Du et al. [44] proposed Recurrent Marked Temporal Point Process (RMTPP) that jointly models the event timings and the event types to forecast the next event type and event timestamp. They use RNN as the building block, and the event marker and its occurrence time are the input for each time step. The idea is to consider the intensity function of a temporal point process as a nonlinear function of the history, and use an RNN to summarize a representation from the event history. Based on this representation, the model predicts the next event type, and event time with an intensity function. Similarly, Xiao et al. [45] presented a model namely recurrent point process networks to predict next event type and its occurrence time. The architecture includes two RNNs; one RNN learns the temporal dependencies among the events (event sequence RNN), and the other RNN models the temporal dependency among the time series (time series RNN). The latent information from both RNNs are fused in the synergic layer by concatenation. Also, an attention mechanism is incorporated into the framework which allows to interpret the strength of previous events on predicting the next event.

Furthermore, Gao et al. [46] proposed a deep learning method for forecasting event subtype. The model considers event subtype forecasting for different locations as a multi-task learning problem. It learns a event subtype representations across tasks through neural networks. The model is based on the assumption that spatially closed locations exhibit similar event subtype patterns. This works also suggests an algorithm based on alternating direction method by multipliers (ADMM) for the optimization of the complex model to solve it more efficiently. Wang et al. [47] forecasted future crime occurrences from past crime data and external data including weather and holiday information. They represented past data through three resolution namely trend, period and nearby. For each of these, the latent representation is obtained through a hierarchical network including convolutional and residual layers. Finally, latent representations are fused together with the external features, and future crime occurrences are predicted. In this study, the input data represented as grid data and spatio-temporal characteristics are learned through convolutional layers. Moreover, Huang et al. [48] proposed a multi-view, multi-modal spatio-temporal learning framework for city-wide abnormal event forecasting. For each of views (event categories, spatial views and semantic views), the model includes an RNN which models the temporal dynamics of corresponding view. Then, an attention layer is applied on top of the latent representations of different views at the same time instant, to automatically learn the importance of view-specific representations. Finally, a conclusive RNN is applied to forecast the abnormal events at the target region. The proposed framework was evaluated on forecasting urban crime and urban anomaly events. Similarly, Huang et al. [49] proposed an attentive hierarchical recurrent network for crime prediction, called DeepCrime. This proposed method considers POIs and other anomalies in the regions to predict crime occurrences for a given region. It uses region embeddings and crime embeddings for each time unit in the past and finds temporal inter-relations with anomalies in the regions using Gated Recurrent Units (GRUs). It also incorporates a temporal attention mechanism to interpret the important time units which are predictive for the occurrences of future crimes. In an another study for crime occurrence prediction, Kang et al. [50] proposed a framework which utilizes DNNs to predict crime occurrences in the city of Chicago. They used multi-modal data obtained from various data sources including image, demographics, education and weather. The proposed model groups spatial, temporal and environmental features, then learns their latent representations through DNNs. Finally, all representations are fused in a joint layer in the deep architecture in order to predict the crime occurrence. Lastly, Jain et al. [51] combine high-level spatio-temporal graphs and RNNs for spatio-temporal modeling for generic tasks including human motion modeling and forecasting. The nodes of the graph typically represent the problem components, and the edges capture their spatio-temporal interactions. Although this method improves the lack of high-level and intuitive spatio-temporal structure of RNNs, it requires an expert knowledge about the domain of the application to identify the problem structure and its decomposition.

2.2.1.2 Time-series Prediction/Forecasting

In addition to event prediction/forecasting studies, there have been a number of temporal/spatio-temporal time series prediction/forecasting works, which have been applied in different domains including air quality prediction [52], wind speed prediction [53], traffic flow prediction [54], vegetation index prediction [55]. Although the purpose of these studies is different from the event prediction task, they treat the modeling/predicting spatial and temporal dependencies problem in a similar manner. Among them, Li et al. [52] proposed a model for predict air quality. They employed stacked auto encoders (SAEs) to extract inherent air quality features and a logistic regression layer on top of it for the prediction. The model was trained in a greedy layer-wise manner. Although it predicts the air quality of all stations simultaneously, it lacks of modeling temporal history of the stations. Fan et al. [56] also introduced a framework based on deep RNN for spatio-temporal air pollution prediction. This framework first fixes missing values by various algorithms. Then, fixed time series are given as input to the deep RNN architecture. To predict the air pollution in the target region, its features and the features from its spatial neighbors are concatenated, and they are fed to the LSTM RNNs. This method ignores the effect of other regions, which are not immediate spatial neighbors of the target region, during prediction and only takes several features of neighbor regions into consideration. In an another study, Bui et al. [57] utilized an LSTM encoder-decoder architecture to forecast air pollution in South Korea. This study showed that using multiple RNN layers in the architecture results in an important increase in the forecasting performance. Furthermore, Ghaderi et al. [53] proposed spatio-temporal forecasting method, which is applied on wind speed forecasting. The proposed method offers separate LSTM models for each time unit in the time window. In other words, each LSTM block is responsible for predicting the wind speed for all stations at the same time for the corresponding time step. The input for the LSTM is the wind speed measurement of all stations for the current time step. In an another study, Das et al. [55] proposed a spatio-temporal deep architecture for vegetation index prediction from remote sensing data, called Deep-STEP. For a given time instance, each pixel is represented by the combination of its vegetation value and that of its neighbors. The architecture consists of stacked MLP layers where each of these is responsible for modeling the previous specific time instance. The output of each layer is concatenated with the input of the next layer. In other words, the temporal evolution is modeled by DNNs irrespective of the spatial dimension.

Moreover, traffic-related prediction problems (e.g., cellular network load prediction, traffic flow prediction, crowd-flow prediction, passenger demand prediction) have been also considered as spatio-temporal modeling/forecasting problem in the literature. Among these studies, Wang et al. [58] proposed a spatio-temporal deep cellular network prediction architecture by combining SAE and LSTM networks. The spatial dimension is represented as grid data. The model employs two types of SAEs, namely local SAEs and a global SAE. For spatial modeling, the architecture uses a separate SAE for each location-of-interest to model local spatial information, and a global SAE to model global spatial information. Then, the learned representations are concatenated and fed to the LSTM. The purpose of the LSTM is to learn the temporal dependencies between different time steps. Lv et al. [59] offered a SAEs based method to predict traffic flows based on the data collected from the detectors. The purpose of SAEs is to extract latent traffic flow features. Each layer of encoder is trained in a sequence so that the output of the previous hidden layer is used as the input of the next hidden layer. On top of SAEs architecture, a logistic regression layer takes places in order to predict the traffic flow. Yu et al. [54] proposed spatio-temporal recurrent convolutional networks for network-wide traffic state forecasting. The network traffic speeds are first converted to the a series of static images. Then, they are given as inputs to the CNN-LSTM architecture. Here CNNs are used to model the spatial dependencies whereas LSTM is used to model the temporal dependencies.

Furthermore, Zhang et al. [9] proposed a deep learning based crowd flow prediction model for spatio-temporal data, called DeepST. The input is grid data (like image) where each cell (region) represents the crowd flow in that cell at a specific time instant. This work describes three different properties (seasonal trend, temporal closeness and period) for the input image. For each property, a separate convolution operation is applied on the temporal dimension, then the outputs are fused. After a number of successive convolutional layers, the future crowd flow (grid or image) is predicted. In an another work, Zhang et al. [60] introduced a spatio-temporal crowd inflow/outflow prediction method, called ST-ResNet. Similar to the prior work, they represent the data as grid data and they have a different component for different temporal influences (closeness, period and trend). Each component includes a CNN and a residual unit. The purpose of these components is to capture the spatial dependencies for different temporal influences. The architecture also consists of an another component to model external features such as day of the week, weather information, etc. The architecture finally fuses outputs of all components and forecasts the future crowd inflow/outflow in each region.

Moreover, Zhou et al. [61] proposed an attention-based encoder-decoder architecture for multi-step city-wide passenger demand. The input for the model is in the shape of grid map. The encoder part consists of convolutional units and ConvLSTM units [62] to capture the spatial and temporal dependencies. The decoder part is the symmetric of the encoder part which consists of ConvLSTMs (copies of the encoder) and deconvolutional layers. The architecture has also an attention component which focuses on the salient parts (annotations) and provides context representation to decoder. The decoder part utilizes the context representation from the attention component and forecasts future passenger demands. Laptev et al. [63] introduced a time-series event forecasting method and applied this method on forecasting the number of completed trips at Uber. Basically, the proposed model consists of two components, namely LSTM autoencoder and LSTM forecaster. The first component consists of stacked LSTM layers in both encoder and decoder parts. The second component contains stacked LSTM layers and a fully connected layer for the prediction. For the prediction, the average of intermediate representations in the LSTM autoencoder is concatenated with the new input, and it is fed to the forecaster component. In addition to aforementioned studies, several studies proposed general purpose time-series forecasting frameworks and applied them in various prediction tasks. Among them, Qin et al. [64] presented a dual-stage attention-based RNN model to make time series prediction, called DA-RNN. It applies input attention to raw data for each time step in order to select most informative series by referring to the previous hidden state. The context information obtained by input attention is fed to LSTM network. Finally, the temporal attention is applied on top of the LSTM network to identify the important time steps which provide more salient contribution to prediction. The proposed model is applied successfully on two different domains (sensory and economics), which are temperature forecasting and index value forecasting of NASDAQ-100. Moreover, Liang et al. [65] introduced an multi-level attentional encoder-decoder architecture for geo-sensory time-series prediction. In the encoder part, it contains two types of spatial attention (as input attentions), namely local spatial attention and global spatial attention. While the former is responsible for identifying important series of target sensor, the latter focuses on the sensors that have more salient contribution on on the prediction. The context information obtained by these two attention is fused and fed to LSTM network. In the decoder part, an another LSTM network with temporal attention take place. The purpose of temporal attention is to select the most informative time steps in the encoder. Note that, the proposed method also incorporates the external features including weather information, sensor ID and time features for the prediction.

Aforementioned DNN-based event and time-series prediction/forecasting studies suffer from two main concerns. First, most of them overlook the complex interactions between local (observed from within a region) and global (observed from all regions) activities across time and space. Only a few have paid attention to this problem, none of the existing works learns to differentiate the pairwise activity relationships between the target location and the other locations. Second, most of the spatio-temporal forecasting studies mainly focus on prediction performance and lack interpretability to uncover the underlying spatio-temporal characteristics of the activities, such as (1) what local and global activity features are more predictive for the subsequent events? (2) what are the locations that have more salient contribution to predicting/forecasting with respect to the target location?

2.2.2 Other Approaches for Temporal/Spatio-Temporal Prediction/Forecasting

2.2.2.1 Point Process-based Methods

One of the mostly used methods for temporal and spatio-temporal modeling is point process. A point process is basically a model of indistinguishable points distributed randomly in some space. Temporal and spatio-temporal point processes [66] have been used across a wide range of domains. Among them, marked temporal point processes have been used in seismology in particularly modeling earthquakes and aftershocks [67, 68]. Short et al. [69] introduced a spatio-temporal point process model for inter-gang violence driven by retaliation in sociology domain (e.g. modeling networks of criminals). Bacry et al. [70] presented a modified version of the non parametric Hawkes kernel estimation procedure in computational finance domain. A Hawkes point process defines random events which are either an immigrant or a descendant. Hawkes process have been also used for various temporal/spatio-temporal domains in univariate or multivariate setting including earthquake analysis [71], violent deaths analysis [72], triggering pattern discovery of spatio-temporal event types [73]. Furthermore, Poisson process [74] and its variants have been employed for human activity modeling [75]. However, typical point process models, including the Hawkes processes [67] and the autoregressive conditional duration processes [76], make specific assumptions about the functional forms of the generative processes, which may or may not reflect the reality, therefore the respective fixed simple parametric representations may restrict the expressive power of these models [44]. Accordingly, for the real applications, the data may be oversimplified or the complexity of the problem may not be captured [77]. In addition to this, the model may underfit due to the misjudgement on model choice.

2.2.3 Markov-based Methods

Markov-based models including Hidden Markov Model (HMM) and Markov Chain models have also been applied in temporal and spatio-temporal modeling and predicting/forecasting domain. Among them, Zhao et al. [7] proposed an enhanced HMM-based model for spatio-temporal event forecasting in Twitter by considering time-evolving context and space-time burstiness. Similarly, Qiao et al. [78] introduced an HMM-based framework that employs temporal burst patterns in an auto-coded event dataset (i.e., Global Data on Events, Location, and Tone (GDELT)) to unveil the underlying event mechanics and predicts the social unrest events by treating it as a sequence classification problem. Alevizos et al. [79] proposed an online system for probabilistic event forecasting based on pattern matching. Given a stream of events, the system forecasts the events using Pattern Markov Chains based on the regular expressions of interested events described by the users. Markov models have also widely been applied for various prediction tasks in the area of energy (e.g., forecasting photovoltaic power generation [80]) and meteorology (e.g., wind power forecasting [81]). Also, Factorizing Personalized Markov Chain (FPMC) has been used for a temporal prediction problem, in particular next basket recommendation [82]. It is a personalized extension of common Markov Chain models, and has become one of the most popular methods for sequential prediction. The main problem of Markov-based models is either that they are based on a strong independence assumption among different factors, or they operate using discrete states and they only take the last known state into consideration.

2.2.4 Tensor Factorization-based Methods

Tensor Factorization (TF) has also been successfully utilized on temporal and spatio-temporal modeling and prediction tasks. For instance, Sehebi et al. [83] offered a TF approach to model learning process of students and predict student performance. It decomposes a tensor, which is created based on the attempt sequences of the students. Bahadori et al. [84] introduced a general TF-based framework for multivariate spatio-temporal analysis. It can incorporate various properties in spatio-temporal data and it can be easily adapted to spatio-temporal forecasting tasks. [85]. Furthermore, Xu et al. [86] presented a spatio-temporal multi-task learning approach, called WISDOM, based on supervised tensor decomposition. It makes predictions based on the aggregation of latent factors obtained by both spatial and temporal prediction models. Similarly, Xu et al. [87] proposed a framework for predictive modeling of multi-scale, spatio-temporal data. It decomposes multiple tensors from different spatial scales so that the latent factors are utilized to train temporal and spatial models. For the prediction, the temporal and spatial outputs of the models are aggregated. For TF-based predictive models, however, it may be hard to predict future behaviors with since it is hard for them to generate latent representations of time bins that have never or seldom appeared in the training data [88].

2.2.4.1 Traditional Statistical Methods

There have been studies on temporal and spatio-temporal event prediction/forecasting which employ traditional statistical methods. Among them, Arias et al. [89] and Bollen et al. [90] used linear regression models with simple features extracted from Twitter to predict the occurrence time of future events. Aghababaei et al. [91] predicted crime rate at a specific city based on the features extracted from Twitter posts. They represented each day as a document-term matrix and predicted the crime rate using past data utilizing linear SVM classifier. Wang et al. [92] applied semantic role labeling (SRL) to Twitter posts and then extracted latent topics using Latent Dirichlet Allocation (LDA) from them for each day. They used latent topics as inputs to forecast the crime ratio via logistic regression. Gerber [93] also employed logistic regression for crime prediction using topics (identified by statistical topic modeling) in social media, particularly Twitter. Korkmaz et al. [94] forcasted civil unrest events using logistic regression with Lasso from online activity data including Twitter and blogs. Similarly, Korolov et al. [95] studied how to predict offline protests from social media. They identified mobilization in social media communication and used this information to predict offline protests via logistic regression. Ramakrishnan et al. [96] proposed an automated system to forecast civil unrests from multiple data sources including Twitter and news outlets. They built several Lasso models for different locations utilizing different types of features (keyword-based, features obtained from follower, retweet-mention graphs) based on the data source. However, aforementioned shallow methods may underfit to model complex interactions among spatial and temporal dimensions to make successful predictions.

2.2.4.2 Advanced Techniques for Prediction/Forecasting

Beside traditional statistical classification and regression models, there have been a number of studies for event prediction/forecasting that employ a variety of advanced techniques including, influence cascade modeling [97], anomaly detection [98], multiple instance learning [2], multi-resolution learning [10, 99], multi-source learning [100], and multi-task learning [101]. Cadena et al. [97] assumed an activity cascade in social media is a precursor for a future offline protest, and they proposed an event forecasting model that uses notion of activity cascades in Twitter based on follower and retweet-mention graph. Rekatsinas et al. [98] proposed a framework, called SourceSeer, which combines spatio-temporal topic models with source-based anomaly detection methods to forecast emergence and progress of rare outbreaks using news sources. The framework basically analyzes the past data in order to detect spatio-temporal patterns of disease. Then, it makes predictions with the second component for possible future outbreaks. Moreover, Ning et al. [2] suggested a nested framework of multiple instance learning to forecast future events as well as jointly identifying the event precursors from news articles. They used text embeddings to represent the articles. Furthermore, Zhao et al. [8] presented a multi-task learning framework that models forecasting tasks in the related geo-locations concurrently using the social media data. This study aims to employ shared information among the locations to improve the forecasting performance. Zhao et al. [99] proposed a multi-task learning framework for multi-resolution spatial event forecasting considering geographical hierarchies between locations. These studies used keywords-based features (e.g. keywords in the message content) for forecasting tasks. Zhao et al. [100] also suggested a model for hierarchical multi-source feature learning to forecast future events. It incorporates multiple data sources with different geographic levels such as social media and currency exchange for civil unrest forecasting; social media and illness surveillance network for influenza outbreak forecasting. Zhang et al. [102] introduced a model for spatio-temporal forecasting of influenza outbreaks based on a domain-specific mechanistic model and demographic information with the enhancement of information from Twitter. Mechanistic model was initialized with information from Twitter (from geo-tagged and topical tweets) and surveillance data was used (from authorized institutions) for the ground-truth. Furthermore, Zhange et al. [103] proposed a spatio-temporal event forecasting framework using hyper-local pricing data, which assumes that variations in pricing data can be precursors for the future events. The proposed framework adopts a tensor completion technique to learn missing values based on spatial and temporal coherence. It also suggests an algorithm to optimize the proposed method based on ADMM. Lastly, Zhao et al. [104] presented a distant supervision of heterogeneous multi-task learning method for societal event forecasting from social media data which considers multiple language setting. The motivation is that events can be followed by social indicators generated by the users who speak different languages. The proposed model first aims to map multi-lingual heterogeneous features to various latent spaces, then using distant supervision it enforces a similar sparsity pattern across them all. However, most of these existing techniques primarily focus on prediction/forecasting performance rather than interpreting spatio-temporal characteristics of the events. In addition, the potential interactions between temporal and spatial dimensions are often overlooked.

2.3 ML Basics for Temporal/Spatio-Temporal Modeling

In this section, the fundamental building blocks for NN-based temporal/spatio-temporal modeling, which are utilized in the proposed approaches, are overviewed. First the RNN structure for sequence/temporal modeling is reviewed. Then, the basics of LSTM which mitigates the drawbacks of traditional RNNs is introduced. After that, attention mechanism which focuses

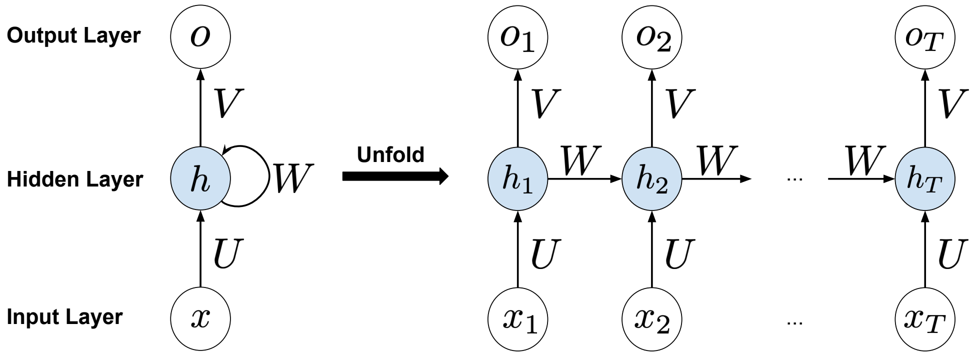


Figure 2.1: Abstract RNN (left) and RNN unfolded over time (right).

on the inputs having more salient contribution for the prediction task is presented. Finally, the information about Group Lasso for neural networks for the selection of important features are provided.

2.3.1 Recurrent Neural Networks (RNNs)

An RNN is basically a densely connected neural network which introduces “time” or “sequence” concept to the traditional neural networks. RNN resolves two main problems of conventional sequence modeling methods, which are modeling arbitrary length sequences and modeling sequences without making any assumptions about the data. RNNs allow for modeling sequences with arbitrary length without making any assumptions about the structure of the data and without considering Markov property [105].

Let x be the sequence of observations (e.g. time series, sentence) where $x = (x_1, x_2, \ldots, x_T)$, T is the sequence length, and x_i is the i^{th} vector representation of input with a dimension of d ($x_i \in \mathbb{R}^d$) for the given sequence x. Figure 2.1 shows the views of a simple RNN which models sequence x. While the left part of the figure indicates the unfolded version of RNN, the right part shows the RNN unfolded over time.

An RNN basically contains an internal memory (i.e. hidden representation, hidden state) h_t, a recursive function f that operates on the input and the internal memory h_t (which is usually a non-linear function), and usually an output function g to produce output (e.g. conditional probabilities) based on the internal memory at time t. The purpose of the internal memory at time t is to summarize the history from x_1 up to x_t. Note that the initial memory h_0 is usually set to zero. The hidden representation at time t is calculated based on the historic summary h_{t-1}, current input x_t and parametric recursive faction f as follows:

$$h_t = f(x_t, h_{t-1}; \theta), \tag{2.1}$$

where $\theta = \{U, W\}$ is the parameter set to be learned in the network. Following the equation above, one of the basic transformation of input and historic summary to current internal memory through recursive function f can be defined as follows:

$$h_t = \sigma(Ux_t + Wh_{t-1}), \tag{2.2}$$

where $U \in \mathbb{R}^{d_h \times d}$, $W \in \mathbb{R}^{d_h \times d_h}$, and d_h is the dimension of the internal memory. The given parameterization is known as Elman Network [106], also called vanilla-RNN in the literature. Once having the internal memory h_t and defining recursive function f, an RNN can also produce an output for each time step through the function g as follows:

$$o_t = g(h_t; V), \tag{2.3}$$

where $V \in \mathbb{R}^{d_h \times d_o}$ to be learned, d_o is the dimensionality of the output o_t. Depending on the task, after the transformation of hidden state h_t by V through the output function g, any function can be applied such as Softmax activation function [107] to output conditional probabilities for the classification task. To sum up, according to Figure 2.1, the input sequence is fed to the RNN network (one observation per time) with transformation matrix U in the input layer; the recursive function f computes the hidden state at the current time step based on the transformed input and previous state in the hidden layer; finally, the output is produced with the output function g by transformation of current hidden state by the V matrix.

Back Propagation Through Time (BPTT) To train an RNN, the gradients of the loss function with respect to the model parameters $\{U, W, V\}$ are needed to be estimated using BPTT algorithm [108, 109]. More specifically, given a loss function $\mathcal{L}(o, y)$, the gradients with respect to the model parameters as $\{\frac{\partial \mathcal{L}}{\partial U}, \frac{\partial \mathcal{L}}{\partial W}, \frac{\partial \mathcal{L}}{\partial V}\}$ are first calculated. Note that the calculation of corresponding gradients at only a specific time t is shown here so that all gradients obtained for each time step should be aggregated at the end. Therefore, the loss for the only time t as $\mathcal{L}_t(o, y)$ is represented. Similar to the typical backpropagation algorithm, BPTT is applied based on the chain rule as well. Accordingly, the gradient of loss function with respect to the parameter matrix V is calculated at time t assuming that no further transformation function applied after the output function g, as follows:

$$\frac{\partial \mathcal{L}_t}{\partial V} = \frac{\partial \mathcal{L}_t}{\partial o_t}\frac{\partial o_t}{\partial V}. \tag{2.4}$$

Furthermore, calculation of the gradients with respect to $\frac{\partial \mathcal{L}}{\partial U}$ and $\frac{\partial \mathcal{L}}{\partial W}$ depends on also the previous time steps (due to the recurrence) in addition to current time step unlike the $\frac{\partial \mathcal{L}}{\partial V}$. The gradient calculation of loss function with respect W at time t is presented as follows:

$$\frac{\partial \mathcal{L}_t}{\partial W} = \frac{\partial \mathcal{L}_t}{\partial o_t}\frac{\partial o_t}{\partial h_t}\frac{\partial h_t}{\partial W}. \tag{2.5}$$

Since the calculation of hidden representation h_t depends on the previous hidden representation h_{t-1} and the corresponding matrix W is employed for the calculation at all time steps,

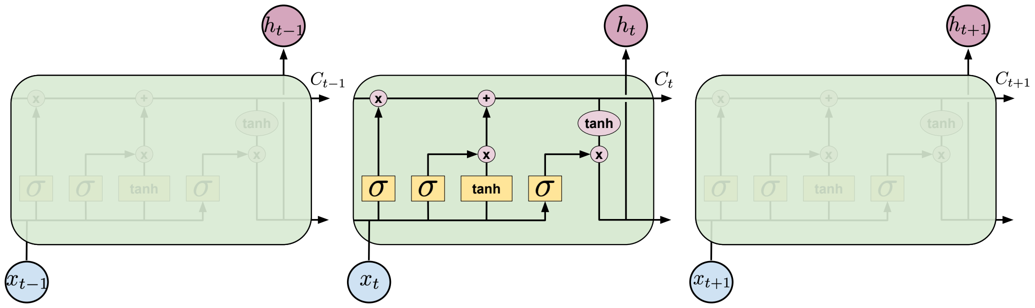


Figure 2.2: Sequential processing in LSTM. In a cell, the pink circles represent the point-wise operations (i.e. addition by $\oplus$, multiplication by $\otimes$) whereas yellow boxes indicate the neural network layers.

the gradients starting from time t to $t = 0$ are needed to be backpropagated, as follows:

$$\frac{\partial \mathcal{L}_t}{\partial W} = \frac{\partial \mathcal{L}_t}{\partial o_t} \frac{\partial o_t}{\partial h_t} \left(\prod_{i=0}^{t-1} \frac{\partial h_{i+1}}{\partial h_i} \right) \frac{\partial h_0}{\partial W}. \tag{2.6}$$

The gradient of the loss function with respect to the matrix U at time t can be also calculated in a similar way. Therefore, only the calculation of $\dfrac{\partial \mathcal{L}_t}{\partial W}$ is provided to be succinct. The total gradients with respect to a specific parameter is calculated by summing the gradients at each time step. For instance, it can be calculated for the matrix W as $\dfrac{\partial \mathcal{L}}{\partial W} = \sum_t \dfrac{\partial \mathcal{L}_t}{\partial W}$. Finally, once the total gradients are calculated with respect to a parameter, the corresponding parameter is updated. The whole process can be done following a training algorithm such as stochastic gradient descent (SGD) [110].

Long Short-Term Memory (LSTM) networks are special type of RNNs, which are able to model long-term temporal dependencies. LSTM was proposed by Hochreiter et al. [111], and it addresses the main problem of basic RNNs, which is exploding and vanishing gradients [112] by using explicit gating mechanisms (input, output and forget gates) to regulate the memory updates. LSTM has been shown to be effective in capturing potential temporal dependencies for different tasks and for a variety of application domains such as text mining [113], diffusion path prediction in social networks [114], video captioning [115].

LSTM cells are capable of storing, writing and reading information through the gates inside the cell. Figure 2.2 indicates the internal representation (gating mechanism) of a simple LSTM cell and unfolded LSTM network. The equations below calculate the information flow within an LSTM cell.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{2.7}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{2.8}$$

$$\widetilde{C}_t = tanh(W_C[h_{t-1}, x_t] + b_C) \tag{2.9}$$

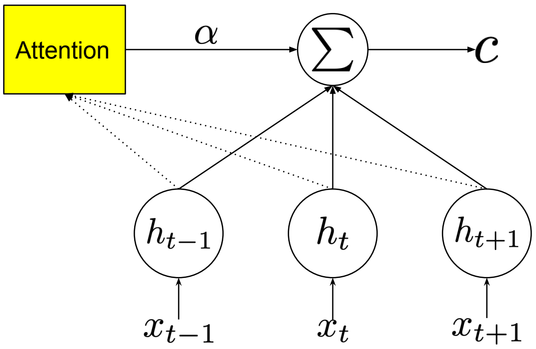


Figure 2.3: An overview of the attention mechanism.

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{2.10}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{2.11}$$

$$h_t = o_t * tanh(C_t) \tag{2.12}$$

where C_t is the cell state at time t, W_f, W_i, W_C, W_o, b_f, b_i, b_C and b_o are the parameters to be learned. The decision of how much information is kept or forgot from the previous cell state C_{t-1}, is performed by forget gate layer (f_t). On the other hand, which values are updated in the current cell is decided by the input gate layer (i_t) meanwhile the new values are computed as $\widetilde{C}_t$. By combining new values and information from the previous cell state, the current cell state C_t is calculated. Finally, what part of the cell state will be output (hidden representation) is calculated based on the current cell state C_t and output gate layer (o_t).

2.3.2 Attention Mechanism

In sequential/temporal modeling, traditional RNNs including LSTM networks summarize the data into a single and fixed-length context vector. However, it can be difficult to cope with the longer sequences for the neural networks. In other words, the earlier parts of the sequence may be forgotten once the entire sequence is processed. Instead of compressing the input into a single, fixed-length context vector, attention mechanism was proposed [116] to take all input representations into consideration and let the network pay more or less attention to each individual input representation while computing the context vector. Attention mechanism has been successfully applied in various tasks including NMT [117], text classification [118], object detection [119].

One possible use of the attention mechanism is to summarize the information from inputs by focusing on the ones that have more salient contribution for the task. For instance; Yang et al. [118] employed such a mechanism for the document classification task where the important words and sentences for the classification task contribute more to the context vector through the attention mechanism. A typical attention mechanism consists of a score function that produces a score for the representation of each input in the sequence. Each input representation contributes to the context vector with respect to its score. The higher score for an input means its more contribution in the context vector. Figure 2.3 simply gives an overview of the attention mechanism. The context vector c is calculated as follows:

$$e_i = v^{\intercal} tanh(Wx_i + b) \tag{2.13}$$

$$\alpha_i = \frac{exp(e_i)}{\sum_{t \in T} exp(e_t)} \tag{2.14}$$

$$c = \sum_{t \in T} \alpha_t x_t \tag{2.15}$$

where $W \in \mathbb{R}^{d \times d}$, $v \in \mathbb{R}^d$, $b \in \mathbb{R}^d$ are the parameters to be learned, e_i is the score for the input x_i, which is calculated by learning another neural network. α_i is the attention weight for the input x_i, In the Eq. (2.14) the attention weights are calculated by soft attention approach using Softmax activation function (to normalize all scores). Finally, the context vector c is computed by taking weighted sum of the inputs with respect to the corresponding attention weights.

Furthermore, attention mechanism has been also incorporated into encoder-decoder architectures and applied for various tasks such as NMT [120], multi-step prediction/forecasting [65], etc. In such architectures, to predict the next output of the decoder, the current state of the decoder is utilized to calculate the contributions of inputs in the encoder. In other words, the current state of the decoder is employed as a *query* and the encoder states are the *keys*. The score function takes the *query* (q) and the *key* (k) as the input. In the literature, several score functions has been proposed. Among them, Bahdanu et al. [116] proposed a *multi-layer perceptron* approach to calculate the score as $score(q, k) = v^{\intercal} tanh(W[q; k])$. Luong et al. [117] proposed a *bilinear* function to calculate score as $score(q, k) = q^{\intercal} Wk$. They also suggested another score function (called *dot product*), which does not need any parameters to be learned but requires that the query and the key should be the same size: $score(q, k) = q^{\intercal} k$. However, the problem of this approach is that the scale of dot product increases as dimensions of the vectors get larger. To solve this problem, Vaswani et al. [121] introduced *scaled dot product* score function which scales the score by the size of the *key* vector as follows: $score(q, k) = \dfrac{q^{\intercal} k}{\sqrt{k}}$. For this kind of attention mechanism, the Eq. (2.13) is replaced with one of the aforementioned score functions.

2.3.3 Group Lasso for Neural Networks

Feature selection is an important step for most of the machine learning tasks dealing with high-dimensional data. To reduce the input dimension, a small subset of input features are searched that brings large amount of discriminative information [122]. From the neural network perspective, the feature selection can be achieved through conventional ℓ_1 regularization in a principled indirect way while optimizing the network parameters at the same time. Basically, ℓ_1 penalizes the sum of absolute values of the weights during training.The ℓ_1 norm acts as a convex proxy of the non-convex, non-differentiable ℓ_0 norm [123]. This penalization can

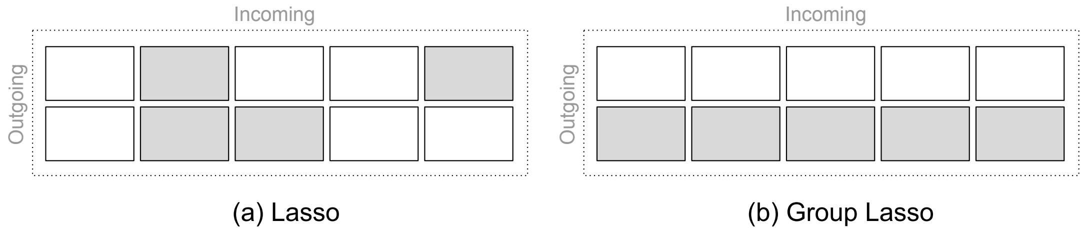


Figure 2.4: (a) Lasso and (b) Group Lasso applied to a single weight matrix. The dark rectangles represent the weights that are set to 0 by the corresponding regularizations (borrowed from [124]).

be defined as follows:

$$R_{\ell_1}(\mathbf{w}) \triangleq \sum_{i=1}^{Q} |w_i|. \tag{2.16}$$

where Q is the total number of weights in the input-level and w_i is the i^{th} input weight. Although ℓ_1 penalization is a principled way to perform sparsity in weight-level, it does not guarantee a structured neural network. In other words, since an input neuron has multiple outgoing weights to the hidden layer, traditional ℓ_1 penalization does not ensure that all the outgoing weights from that input neuron will be 0. To achieve group-level sparsity (neuron-level) as well as optimization of the network at the same time, Scardapane et al. [124] proposed a group sparse regularization method for deep neural networks. This regularization imposes sparsity on a group level, such that all the weights in a group are either simultaneously set to 0, or none of them are. The Figure 2.4 shows an example comparison between plain ℓ_1 regularization and group lasso regularization on the input-level of a simple neural network. There are two input neurons, and there exist five neurons in the hidden layer. While Lasso sets some of the input weights to 0 of both input neurons, Group Lasso provides sparsity in the group level. According to the figure at the right, the first input can be removed from the network since its all outgoing weights are set to 0 by Group Lasso.

Group Lasso has been shown to be effective in several domains, such as robotic control [125] and multi-modal context [126] to select informative features. It can be used to interpret the neural network model in such a way that redundant information from features are minimized, which allows for differentiating which features are important for a prediction task. For feature selection in a neural model, Group Lasso regularization can be formulated as follows [127]:

$$R_{\ell_{2,1}}(\mathbf{w}) \triangleq \sum_{g \in G} \sqrt{|g|}\|g\|_2 \tag{2.17}$$

where g is the vector of outgoing connections (weights) of an input neuron, G denotes a set of vectors of the input neurons, and $|g|$ indicates the dimension of vector g. $\|.\|$ is the Frobenius norm.

2.4 Chapter Summary

In this chapter, a background information about several population-level societal events including social movements and opioid overdoses was given. The possible links between human behavior and these societal events based on the relevant social theories were presented. It was shown that given human behaviors can be employed for predicting the future societal events. Next, the recent approaches for temporal/spatio-temporal predicting/forecasting were reviewed. These approaches were divided into two main categories, namely DNN-based approaches and other approaches. Based on the findings from the literature, these approaches suffer from two main drawbacks. First, most of these studies overlook the complex interactions between temporal and spatial dimensions while modeling. Also, the relationships between local (observed from within a region) and global (observed from all regions) activities across time and space are mostly disregarded. Second, most of the spatio-temporal predicting/forecasting studies primarily concentrate on prediction performance and provide none or limited interpretability to unveil the underlying spatio-temporal characteristics of the activities. Accordingly, it is important to consider and seek to find solutions to these problems in the proposed methods. Finally, in the last part, a background information about basic machine learning concepts that are employed in the proposed methods were provided within the scope of this thesis.

CHAPTER 3

ACTATTN: A NOVEL DEEP SPATIO-TEMPORAL METHOD FOR MODELING AND FORECASTING SOCIETAL EVENTS

This chapter proposes a novel deep spatio-temporal predictive model to forecast societal events, called ActAttn. Using spatio-temporal data, it seeks to characterize the social, spatial, and temporal features in relation to the subsequent societal events in a unified and automatic manner. For this purpose, a deep learning architecture is developed, which is not only capable of forecasting the occurrence of future protests, but also allows for interpreting what features, from which places, have significant contributions on the protest forecasting model, as well as how they make those contributions. To accomplish this, a two-level attentional network architecture is introduced, which (a) differentiates the feature contribution from local (intra-region) and global (inter-region), and (b) identifies the regions, referred as the “hubs”, that have a more salient contribution in predicting protest events globally.

The proposed model is evaluated on the domain of social movements, where the purpose is to forecast future offline protests from spatio-temporal social media data. The diffusion of new information and communication technologies, social media (e.g. Twitter) has been central in providing mobilizing information, coordinating demonstrations, and creating opportunities for people to exchange opinions [12, 13]. The lexicon approach is utilized to extract a range of theory-driven linguistic features that allows for making sense of the association between the types of activity traces in social media and future offline protests. The proposed model ActAttn is validated on three social movements datasets through extensive experiments. Also, in-depth analysis and comparison across several baselines and state-of-the-art methods are performed.

This chapter first defines the forecasting problem on a specific societal event domain (i.e. forecasting offline protests), where the proposed method is applied and evaluated. Next, the architecture of the proposed model is presented in detail. Then, the experiment details are given including datasets, extracted features, comparison methods and experimental settings. After that, the experiment results are provided including in-depth analysis and comparison across several baseline and state-of-the-art methods. Finally, the discussion and conclusion about the proposed method are presented including the limitations of the current work.

3.1 Problem Definition

Suppose there are L locations (e.g., cities, states) of interest, and each location l can be represented by a collection of static and dynamic features. The static features (e.g. population, political leaning) are features that remain the same or change slowly over a longer period of time, and the dynamic features (e.g., percentage of tweets that express the "anger" emotion) are updated for each time interval t (e.g., hour, day). Let S_l be the set of static features of location l, and $X_{t,l}$ the set of dynamic features for location l at time t. There is also given a binary variable $Y_{t^*,l} \in \{0, 1\}$ that indicates the occurrence of a future protest event for each location l at time t^*. The collection of dynamic features from all locations within an observing *time window* with size k up to time t can be represented as $\mathcal{X}_{t-k+1:t} = \{\mathcal{X}_{t-k+1}, \ldots, \mathcal{X}_t\}$, where $\mathcal{X}_{t'} = \{X_{t',1}, \ldots, X_{t',L}\}$.

The purpose is to predict the future event occurrence $Y_{t^*,l}$ at specific location l at a future time $t^* = t + \tau$, where τ is called the *lead time* for forecasting. The forecasting is based on the static and dynamic features of the location itself, as well as the dynamic features in the environment (from all other locations). Therefore, the forecasting problem can be formulated as learning a function $f(S_d, \mathcal{X}_{t-k+1:t}) \rightarrow Y_{t^*,d}$ that maps the input, the static and dynamic features, to a protest indicator at the future time t^* for a *target* location d.

In order to facilitate interpretation of the protest forecasting, it is sought to develop a model that can differentiate the contribution of the features, the locality (local/intra-region features vs. global/inter-region features), and the overall importance of each location when contributing to the prediction of other locations. Therefore, the dynamic features $\mathcal{X}_{t-k+1:t}$ are further organized into two sets: the *intra-region* features, $\{X_{t-k+1,d}\ \ldots, X_{t,d}\}$ represent the sequence of dynamic features for the location d, and the *inter-region* features, $\{X_{t-k+1,l}\ \ldots, X_{t,l}\}$ for $l \in \{1, 2, \ldots, L\}$, contain the sequences of dynamic features for all locations of interest.

3.2 Proposed Architecture

As shown in Figure 3.1, the proposed architecture involves three primary components, the local component $\mathcal{M}^{loc}$, the global component $\mathcal{M}^{glob}$, and the static features S_d. S_d provides location-specific information about the target location d. The (intra-) local component $\mathcal{M}^{loc}$ is designed to model the contribution of the local dynamic features (*intra-region* features) for the target location. The global component $\mathcal{M}^{glob}$ is to model the spatio-temporal contribution of dynamic features for all locations of interest (*inter-region* features). The input for the (intra-) local component is $\{X_{t-k+1,d}\ \ldots, X_{t,d}\}$ while the input for the global component is $\{X_{t-k+1,l}\ \ldots, X_{t,l}\}$ for $l \in \{1, 2, \ldots, L\}$.

The Recurrent Unit. In both $\mathcal{M}^{loc}$ and $\mathcal{M}^{glob}$, LSTM is used as a building block in the proposed model to capture the temporal relationships among the dynamic features. LSTM has been shown effective in capturing potential temporal dependencies, and it addresses the vanishing and exploding gradient problems of basic RNNs by using explicit gating mechanisms (input, output and forget gates) to regulate the memory updates. A single LSTM network is included to model intra-region dynamics in $\mathcal{M}^{loc}$ (Figure 3.1-c). To capture the spatio-temporal relationships among all locations in $\mathcal{M}^{glob}$ (Figure 3.1-b), separate local components are in-

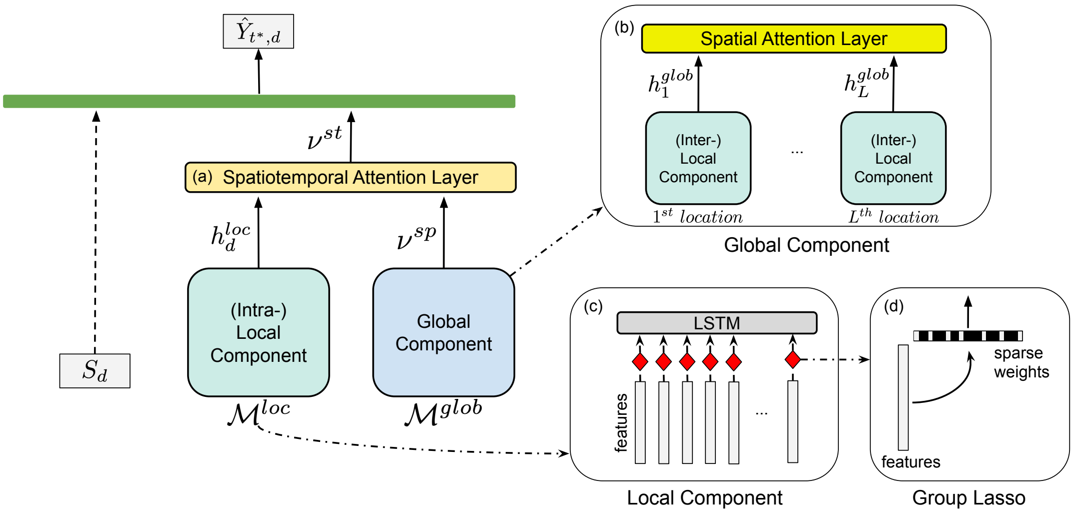


Figure 3.1: Overview of our proposed ActAttn architecture [128]. It incorporates hierarchical attentional networks where the top level (a) differentiates the intra-region and inter-region importance, the second level (b) identifies the hub regions. The temporal dependency of time-varying features in both intra- and inter-regions are modeled using LSTM (c), with sparse feature learning using Group Lasso regularization (d).

cluded, called (inter-) local components, where each of them has the same structure as $\mathcal{M}^{loc}$. Each (inter-) local component is then responsible for modeling the temporal dynamics of a single location. The LSTM outputs (hidden states) inside $\mathcal{M}^{loc}$ and $\mathcal{M}^{glob}$ are h_d^{loc} and $\{h_1^{glob}, h_2^{glob}, \ldots, h_L^{glob}\}$, respectively. Note that h_d^{loc} and any h_l^{glob} (where $l \in \{1, \ldots, L\}$) are used for the last hidden state of the corresponding LSTMs (at time t) shortly. They are calculated as follows:

$$h_d^{loc} = f(h_{d,t-1}^{loc}, X_{t,d}) \tag{3.1}$$

$$h_l^{glob} = s_l(h_{l,t-1}^{glob}, X_{t,l}) \tag{3.2}$$

where $f(.)$ and $s_l(.)$ are the LSTMs in the (intra-) local component and the (inter-) local component for the location l.

Hierarchical Attention Mechanism. Attention mechanism has been found powerful in reweighting the internal components in a neural architecture [116, 129]. A hierarchical attention mechanism is designed to differentiate the importance of spatial and temporal information. First, in $\mathcal{M}^{glob}$, a spatial attention layer is incorporated on top of $\{h_1^{glob}, h_2^{glob}, \ldots, h_L^{glob}\}$ to learn the spatial importance among all locations (Figure 3.1-b). The idea is that not all the locations contribute equally to the prediction of event occurrence at a target location, and this attention layer is to reward the locations which contribute the most to forecast the event occurrence in the target locations. The context vector ν^{sp}, which is the output of the global component, is computed as follows:

$$e_i = (v^{sp})^{\intercal} tanh(W^{sp} h_i^{glob} + b^{sp}), \tag{3.3}$$

$$\alpha_i = \frac{exp(e_i)}{\sum_{l=1}^{L} exp(e_l)}, \tag{3.4}$$

$$\nu^{sp} = \sum_{l=1}^{L} \alpha_l h_l^{glob}, \tag{3.5}$$

where $W^{sp} \in \mathbb{R}^{m \times m}$, $b^{sp} \in \mathbb{R}^m$ and $v^{sp} \in \mathbb{R}^m$ are the parameters to be learned, m is the number of hidden units in LSTMs in each (inter-) local component. ν^{sp} is the spatial attention layer output (the context vector) that summarizes the aggregate contribution of all locations, α_i is the attention weight for the location i. Second, a spatiotemporal attention layer is introduced to differentiate the local (intra-region) and the global (inter-region) feature contributions (Figure 3.1-a). The idea behind this layer is that, in some cases, the occurrence of societal events may largely depend on the temporal information within the locations themselves, while in other cases, the occurrence may depend more on the happenings of other locations or the global dynamics. The spatiotemporal attention layer is given by:

$$u_{loc} = (v^{st})^{\intercal} tanh(W^{st} h_d^{loc} + b^{st}), \quad u_{glob} = (v^{st})^{\intercal} tanh(W^{st} \nu^{sp} + b^{st}), \tag{3.6}$$

$$\beta_{loc} = \frac{exp(u_{loc})}{exp(u_{loc}) + exp(u_{glob})}, \quad \beta_{glob} = \frac{exp(u_{glob})}{exp(u_{loc}) + exp(u_{glob})}, \tag{3.7}$$

$$\nu^{st} = \beta_{loc} h_d^{loc} + \beta_{glob} \nu^{sp}, \tag{3.8}$$

where $W^{st} \in \mathbb{R}^{m \times m}$, $b^{st} \in \mathbb{R}^m$ and $v^{st} \in \mathbb{R}^m$ are the parameters to be learned, m is the number of hidden units in LSTMs (inter-) and (intra-) local components. β_{loc} and β_{glob} are the attention weights corresponding to the outputs of (intra-) local and global components, respectively. ν^{st} is the spatio-temporal context vector that aggregates the information learned from temporal and spatial dimensions from both local and global components. Finally, the forecasting of the occurrence of an event at future time t^* and at the target location d is computed using spatio-temporal context vector and static features of the target location as follows:

$$\hat{Y}_{t^*,d} = \phi(W_f [S_d, \nu^{st}] + b_f), \tag{3.9}$$

where S_d is the static feature of the target location d, $W_f \in \mathbb{R}^{(m+o) \times (m+o)}$ and $b_f \in \mathbb{R}^{(m+o)}$ are the weight matrix and bias vector to be learned in the final concatenation layer, respectively. m is the number of hidden units in LSTMs (inter-) and (intra-) local components whereas o is the static feature size. ϕ is the activation function where we apply *Softmax* activation function in order to obtain posterior probabilities of occurrence and non-occurrence of the event.

Objective Function. The objective function is composed of two loss terms, which are prediction loss and Group Lasso regularization. It is formulated as follows:

$$\mathcal{L}_{overall} = \mathcal{L}_{predict} + \lambda \mathcal{L}_{GL}, \tag{3.10}$$

where λ is the regularization coefficient for Group Lasso regularization. $\mathcal{L}_{predict}$ is the cross-entropy loss which penalizes the prediction of the network for the possible future events as follows:

$$\mathcal{L}_{predict} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} Y_{ij}\log(p_{ij}), \tag{3.11}$$

n is the number of samples, c is the number of class labels (occurrence of an event and non-occurrence of an event in our case), and p_{ij} is the probability of the sample i assigned to class j by the model. Furthermore, the Group Lasso regularization is incorporated into the loss function. Group Lasso has been shown effective in several domains such as robotic control [125] and multi-modal context [126] to select informative features. The main motivation of employing this regularization is to select informative features in (intra-) local and (inter-) local components (Figure 3.1-d) while assigning the optimal weights of the network at the same time. Therefore, it also enables us to interpret the model in such a way that redundant information from features are minimized, which allows for differentiating which features are important for the occurrence of future events. The loss term for Group Lasso is defined as follows:

$$\mathcal{L}_{GL} = \lambda_1 \left\| W^{loc} \right\|_{2,1} + \lambda_2 \sum_{l=1}^{L} \left\| W_l^{glob} \right\|_{2,1}, \tag{3.12}$$

W^{loc} is the input weight matrix in $\mathcal{M}^{loc}$, W_l^{glob} is the input weight matrix of (inter-) local component of l^{th} location in $\mathcal{M}^{glob}$. Note that the input weight matrix contains all weights of LSTM except for recurrent and bias weights. Moreover, λ_1 and λ_2 are the regularization factors for $\mathcal{M}^{loc}$ and $\mathcal{M}^{glob}$, respectively. Therefore, each component can be regularized by different regularization factors. The Group Lasso regularization for a given weight matrix W can be written as follows:

$$\|W\|_{2,1} = \sum_{g \in G} \sqrt{|g|}\, \|g\|_2 , \tag{3.13}$$

where g is vector of outgoing connections (weights) from an input neuron, G denotes a set of input neurons, and $|g|$ indicates the dimension of g. Each input neuron in $\mathcal{M}^{loc}$ and in each (inter-) local component of $\mathcal{M}^{glob}$ is represented as a separate group so that G contains vectors of these groups.

3.3 Experiments

This section provides information about the datasets (including data collection process) that are used in the experiments to evaluate our method, the feature sets (including feature extraction) that are employed in our model, the comparison methods for the evaluation, and the experimental settings.

3.3.1 Datasets

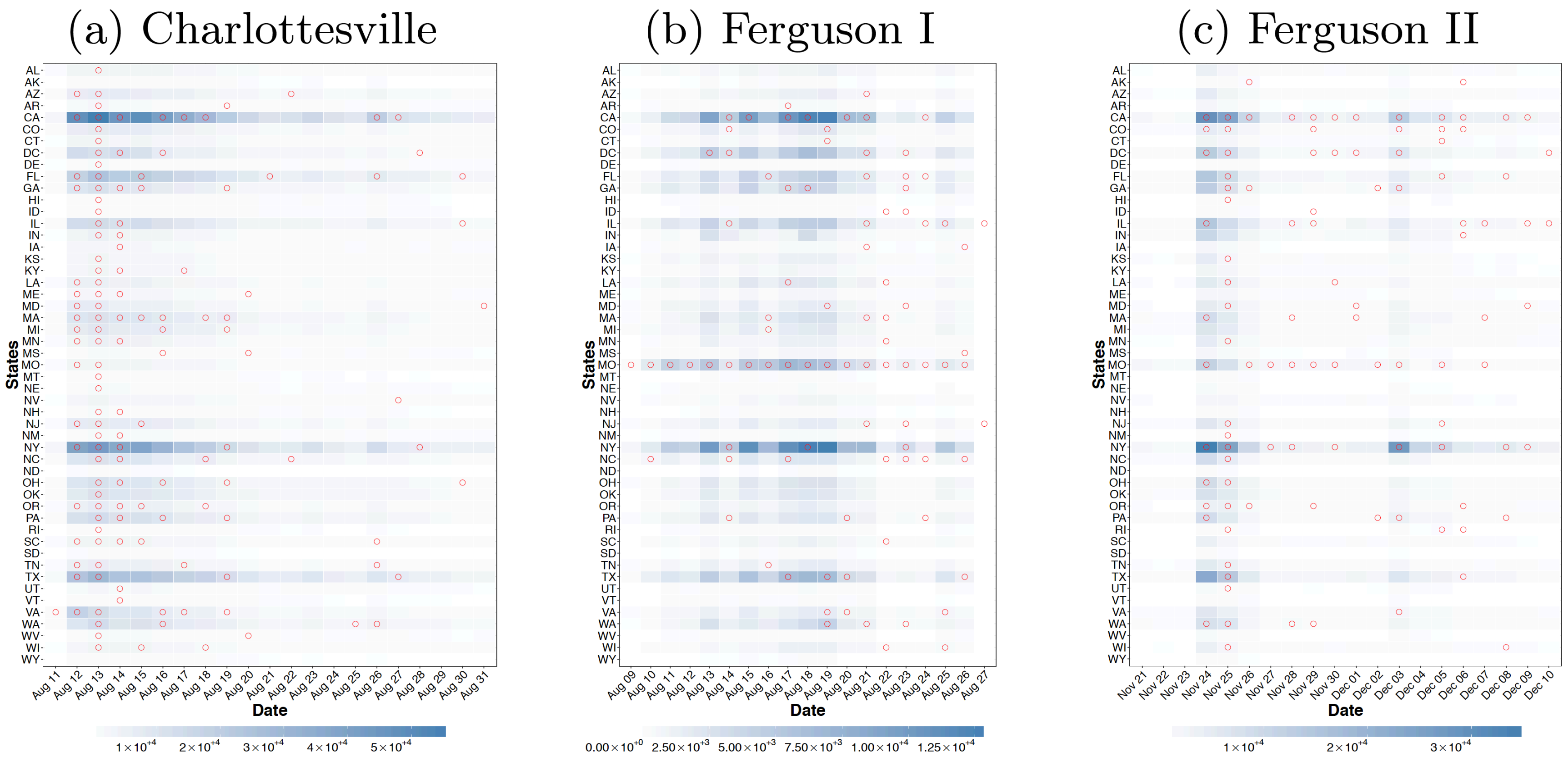
The experiments were performed on three different social movements. For each of them, the same type of datasets were collected. The social movements were chosen with respect to their social significance and meanwhile in a way that they would allow to test the design of the proposed model that could take care of the distinct social, temporal, and spatial dimensions of the nature of protests. Moreover, the nature of issues in the chosen movements should not be too different in order to compare and contrast the performance of the social theory-driven features. Eventually, two social movements were considered, which are Black Lives Matter (BLM) and counter-protests to Charlottesville's white supremacist rally [130]. For BLM, two separate waves of protests (*Ferguson I* and *Ferguson II*) were selected regarding the police's killing of Michael Brown in Ferguson, Missouri [3]. The first wave just started after the police's killing of Micheal Brown whereas the second wave started just after the grand jury decision not to indict Darren Wilson (the police that shot Michael Brown) and then spread across the U.S. Ferguson unrests are symbolic protests in BLM against systemic racism to Black people in the U.S. Charlottesville counter-protests are the largest recent nationwide protest activities against white supremacism in the U.S. As shown in Figure 3.2, these different social movements left heterogeneous activity traces, both online and offline, over time and across locations, creating significant challenges in analyzing their spatial and temporal patterns. For each of the social movements (*Charlottesville*, *Ferguson I* and *Ferguson II*), the related datasets were collected as follows:

3.3.1.1 Twitter Data

Tweets were collected based on specific keywords or hashtags relevant to the counter-protests to Charlottesville rally, and the first and the second waves of Ferguson protests. The size and statistics of each dataset are provided in Table 3.1. ***Charlottesville Dataset*** was collected

Table 3.1: Basic statistics of the Twitter and protest datasets.

Dataset	Duration	#Tweets	#Users	#Protest Occurrences
Charlottesville	Aug 11 - Aug 31 (2017)	11.36M	5.93M	136
Ferguson I	Aug 9 - Aug 27 (2014)	8.02M	2.76M	90
Ferguson II	Nov 21 - Dec 10 (2014)	9.86M	3.80M	104



(a) Charlottesville

(b) Ferguson I

(c) Ferguson II

Figure 3.2: Spatio-temporal occurrence for different social movements by date (x-labels) and by location (y-labels). A red circle indicates at least one offline protest event happening on a particular day and state; the blue shade indicates the volume of tweets posted at the corresponding day/state. Charlottesville counter-protests exhibited burst patterns where most of the activities were sparked by a deadly violence attack and President Trump's statements on Aug 12th, 2017. In the first few days following the attack and the statements, more protest events occurred nation-wide and larger tweet volume was observed. The Ferguson I protests appeared to have a gradual build-up process where the activities are initially local (around Missouri and few states) with the shooting of Michael Brown on Aug 9th, 2014 and later received global attention. Until Aug 20th, 2014, a global increase in tweet volume was observed. The Ferguson II protests started on Nov 24th, 2014 with the announcement of jury decision not to indict the police officer and took global attention. The tweet volume for each state are more in the first two days after jury decision compared to other days.

through the Twitter Streaming API[1] by 17 keywords and/or hashtags of interest[2]. Retweets were not included. These keywords were emerging during the event happenings and were then widely used on Twitter to refer to the relevant issues and happenings. The ***Ferguson I Dataset*** and ***Ferguson II Dataset*** were collected based on the published work [131], using 45 keywords including #ferguson, #blacklivesmatter, "black lives matter" and the names of Black people killed by police during 2014 and 2015. Based on the tweet IDs provided in the aforementioned published dataset, the tweets were recollected within the two periods via Twitter REST API[3] and the retweets were excluded.

3.3.1.2 Protest Data

The ground-truth data was collected from the website of Elephrame[4] on the occurrence of offline protest events during the periods of Charlottesville counter-protests, and the two waves of the Ferguson protests. Elephrame provides information about civil unrest events occurred in the US. This information is kept in a structured way and includes protest occurrence time (start date and end date), protest location (in state-level and city-level), protest subjects (subtype of the protest event), description, number of participants, and at least one source link. The Elephrame website was crawled and the protest information was collected through python libraries *Scrapy*[5] and *Selenium*[6] in December 2017. News reports about BLM protests were also incorporated that were collected by the authors of [19]. Each protest event information is based on the given source link(s). Note that there can be more than one event in the same location at the same time interval. In this work, it was only considered whether a protest event occurred in a given location at that time interval. The occurrence of a protest event was represented using binary variable. As a result, 136, 90 and 104 offline protest events have been observed across the country during the three movements Charlottesville, Ferguson I and Ferguson II, respectively. While the tweets for Charlottesville and two waves of Ferguson were collected separately using different collection methods (Streaming API vs. REST API), the information about protest events was collected from the same data source – the Elephrame website. Since the main focus of this work is the spatio-temporal patterns of the offline protest events, the difference in terms of the methods used for collecting tweets does not significantly impact the results and the interpretation.

3.3.1.3 Census Data

The 2010 United States Census data was used, which is provided by the U.S. Census Bureau, to extract the static features related to demographics (population, population density and diversity). The Census data contains varying types of information from demographics to economical indicators for different spatial resolutions.

[1] https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data.html

[2] Keywords include: Charlottesville, KKK, Ku Klux Klan, Klansman, Klansmen, Nazi, Nazism, racism, racist, supremacy, supremacist, supremacists, #Charlottesville, #domesticterrorism, #FireBannon, #WhiteSupremacist, #WhiteSupremacists

[3] https://developer.twitter.com/en/docs/tweets/post-and-engage/api-reference/get-statuses-show-id.html

[4] https://elephrame.com/

[5] https://scrapy.org/

[6] https://selenium-python.readthedocs.io/

3.3.1.4 Location Extraction

In this work, the purpose is to forecast the occurrence of offline protest events at the state level, using Twitter users' activities. The locations of tweets were either extracted from their geocodes (if available) or inferred from the users' profiles. First, the geo-tagged tweets posted from the United States included state information in their 'place' field. These kinds of posts include either a state name or state code (2 letter-code). This kind of information was directly used as the location indicator. Second, the location information of the tweets are extracted from the user profiles. This approach was followed for the tweets whose locations cannot be identified using the first approach. Similar to the first approach, the locations (state name or state code) were identified if they were explicitly written in the user profiles. If they were not, the names of the cities located in the United States were also checked. If a city name could be identified in the profile, it was mapped to its corresponding state. For this purpose, a dictionary was used from Encyclopedia Britannica[7] which includes city-state pairs in the United States. Note that there can be more than one city with the same name in different states. Therefore, such cities were discarded in this study. In total, the tweet locations at the state-level for 29.9%, 41.5% and 43.3% of all tweets in the Charlottesville, Ferguson I, and Ferguson II datasets were able to be extracted, respectively.

3.3.2 Features

As mentioned earlier, two types of features are incorporated into the proposed model, which are the static features and the dynamic features.

Static features reflect political and demographic backgrounds of a location that a protest event may take place, including the location's belonged state's *population*, *population density*, *voting behavior* (we used vote ratio for Trump in the 2016 presidential elections of United States as an indicator of the degree of conservative of the location), and coarser grained *region information that the location belongs to* (i.e. North-east, Mid-west, South and West). These static features either remain unchanged or change slowly over time. For the feature *population*, the log transformation was applied for the normalization purpose.

Dynamic features are to capture social media user's online activities that may be predictive of offline protests. These features change for each time interval for a given location. Drawn upon social movement literature [14] (discussed in Section 2.1), four factors were considered to extract dynamic features, namely *emotion*, *identity*, *grievance*, and *social embeddedness*.

A lexicon-based approach was followed to extract the dynamic features. Three dictionaries (LIWC [132], SentiSense [133], and Moral-Laden [134]) were employed to capture these features indicating *emotions*, *grievance*, and *identity*. These dictionaries consist of categories (e.g. *anger* from LIWC, *disgust* from SentiSense, *Ingroup Virtue* from Moral-Laden), and each category contains words and/or word stems. The categories which indicate *emotion*, *identity*, *grievance* will correspond to features in the feature set. Several additional relevant features beyond these key factors were also included to test their usability. To extract the dynamic features for *emotions*, *grievance*, and *identity*, the tweet contents were first parsed and lemmaization was applied. Then, the words that match with dictionary words or derived

[7] https://www.britannica.com/topic/list-of-cities-and-towns-in-the-United-States-2023068

Table 3.2: Mapping of features to corresponding key factors.

	Emotion	Grievance	Identity	Social Embeddedness	Other
LIWC	posemo, negemo, anx, anger, sad.	negate, verbs.	p1, p2, p3, social, friend, family.	-	discrep, hear, feel, death, swear, past, present, future, article.
SentiSense	All	-	-	-	-
Moral-Laden	-	All	-	-	-
Twitter Engagement	-	-	-	All	-

from dictionary word stems were identified. To do so, for each tweet, it was decided whether it contains any keyword related to each of the given categories. LIWC and SentiSense include a range of emotions, either positive or negative; LIWC offers the categories of *social* and *personal pronouns* that may serve as indicator of identity. On the other hand, Moral-Laden dictionary was used to capture grievance. This dictionary is derived from moral foundation theory that consider humans engage in moral judgments along at least five dimensions: Harm/Care, Cheating/Fairness, Betrayal/Loyalty, Subversion/Authority, and Degradation/Purity. While grievance results from the appraisal of relative deprivation based on moral rules, the Moral-Laden is used with an attempt to capture the grievance. In this work, the categories (features) from LIWC contain *positive emotion (posemo)*, *negative emotion (negemo)*, *anxiety (anx)*, *anger*, *sadness (sad)*, *social process (social)*, *family*, *friends (friend)*, *discrepancy (discrep)*, *hear*, *feel*, *death*, *swear words (swear)*, *past focus (past)*, *present focus (present)*, *future focus (future)*, *common verbs (verbs)*, *articles (article)*, *negations (negate)*, *1st person plural (p1)*, *2nd person (p2)* and *3rd person plural (p3)*. The categories (features) from SentiSense include *love*, *joy*, *fear*, *hate*, *ambiguous*, *anticipation*, *like*, *sadness*, *despair*, *calmness*, *disgust*, *surprise* and *hope*. The categories (features) from Moral-Laden consist of *ingroup vice*, *ingroup virtue*, *harm vice*, *harm virtue*, *fairness vice*, *fairness virtue*, *authority vice*, *authority virtue*, *purity vice*, *purity virtue* and *morality general*.

Furthermore, to identify the type and level of *social embeddedness*, social media users' engagement in online discussion was captured as features, including *number of tweets (num_tweet)*, *number of reply tweets (num_reply)*, and *number of tweets with URL links (num_urlTweet)*. Greater volumes of either type of tweeting behaviors (tweets, replies, and URLs) may reflect that the public could be more aware of focal issues and events. Therefore, they turn to be more motivated in seeking, spreading, and exchanging information, ideas and emotions in social media. Such social contexts may raise individual's perception of the efficacy of protests that could lead to actual protest actions. More replies and URL links suggest being more embedded in relevant social networks. Replies suggest direct interactions with other embedded users. On the other hand, URL links indicate the information networks built based on relevant information/content created by others.

The mapping of the features to the corresponding key factors (*emotion*, *grievance*, *identity*, *social embeddedness* and *other*) are provided in Table 3.2. Note that for each time unit (i.e.

time interval) and for each location, the features were aggregated to create the feature vectors. Each feature based on LIWC, SentiSense and Moral-Laden dictionaries were normalized by the number of tweets. For instance, for the feature *anger* from LIWC, the number of tweets that include information related to that feature for a specific time interval and a specific location was normalized by the number of all topic-related tweets. This score was assigned to *anger* feature for the given time interval and the location as the feature value. Finally, the feature *num_tweet* was applied zero-mean and unit variance normalization.

3.3.3 Comparison Methods

The proposed method ActAttn was compared with several baselines and state-of-the-art approaches. In order to evaluate the forecasting effectiveness of the proposed model, three sets were selected as the comparison methods.

The first set includes Logistic Regression (LR) and Support Vector Machine (SVM) classifiers since they are widely-used machine learning methods in the event detection/forecasting literature. In these methods, the effect of static, intra-region and inter-region features by combining all features together are examined. The second set of methods includes recently developed neural-network based models, such as RNNs and LSTMs in particular, as they have been shown to be a superior performance in event forecasting problems due to the capability of modeling the temporal dependencies. The third set of methods are the state-of-the-art spatio-temporal event forecasting approaches recently proposed by [8], including regularized multi-task feature learning *(RMTFL)*, constrained multi-task feature learning I *(CMTFL-1)* and constrained multi-task feature learning II *(CMTFL-2)*. These methods formulate event forecasting for multiple locations as a multi-task learning problem. They build event forecasting models for different locations simultaneously by restricting all locations to select a common set of features. Note that none of the existing approaches support the hierarchical structure of features coming from intra- and inter-regions, and we will discuss the importance of such differentiation more in the Section 3.4. The baseline methods are summarized as follows:

The first set:

- *Logistic Regression (LR)* is simple LR model. Three baselines were included for this model. $LR[tem]$ uses only intra-region features, $LR[s, tem]$ concatenates static and intra-region features, and $LR[s, tem, st]$ merges all types of features as the input.
- *Support Vector Machine (SVM)* is simple SVM model. $SVM[tem]$ employs only intra-region features while $SVM[s, tem]$ combines static features with intra-region features. Also, all features are employed as the input in $SVM[s, tem, st]$.

The second set:

- *LSTM* is a basic LSTM network that employs only intra-region features. Although it models the temporal dependencies, it does not consider static features and spatial relationships among regions.
- *S + LSTM* is an LSTM-based model where intra-region features are given as inputs to LSTM network. Then, the embeddings of dynamic features (hidden representations) are

concatenated with the static features. Although this model takes the social properties of the locations into consideration, it does not consider the spatial relationships among regions.

- *S + LSTM (GL)* has the same structure as $S + LSTM$, yet it was trained incorporating Group Lasso regularization into loss function. With this model, the purpose is to monitor the effect of Group Lasso regularization on the performance of $S + LSTM$ model.

  The third set:

- $RMTFL$ employs a regularization parameter to control the model sparsity.
- $CMTFL$-1 introduces a constraint to control the number of features in the model for sparsity.
- $CMTFL$-2 restricts the number of features selected from static and dynamic groups separately.

Furthermore, to evaluate the effectiveness of individual components of ActAttn, including the Group Lasso regularization and hierarchical attention mechanism (spatial and spatiotemporal attentions), several variants of ActAttn were included for comparison as follows:

- *ActAttn (w/o GL)* has the same structure of our proposed method, yet Group Lasso regularization is not incorporated into the loss function.
- *ActAttn (w/o stAttn)* does not include the spatiotemporal attention layer instead h_d^{tem} and v^{sp} are concatenated. To do so, the aim is to evaluate the importance of reweighting the contributions from (intra-) local and global components on the forecasting performance.
- *ActAttn (w/o spAttn)* does not include the spatial attention layer, instead a linear projection layer is used. With this variant, the purpose is to observe the effect of reweighting the contributions of all locations in the global component on forecasting.

3.3.4 Experimental Settings

In the experiments, 'day' was used as the time unit and 'state' was used as the location unit. These units were chosen based on availability of the data and short-term nature of protest occurrences. The last 5 days from each dataset were used as the test sets and rest of them as the training sets. The training set of *Charlottesville* dataset contains 127 protest events (15.6% of all samples in the training set) and test set contains 9 events. The training set of *Ferguson I* dataset contains 63 protest events (9% of all samples in the training set) and test set contains 27 events. The training set of *Ferguson II* dataset contains 82 protest events (10.7% of all samples in the training set) and test set contains 22 events. Different settings of window size and lead time were applied. The window size k was set to be $\{1, 2, 3\}$ and the lead time τ was set to be $\{1, 2, 3\}$. The hidden unit size for LSTM is 16. The architecture was trained using the Adam optimizer [135] with a learning rate of 0.001. For the models incorporating Group Lasso regularization, regularization factors λ_1 and λ_2 were selected from the set $\{10^{-5}, 10^{-4}\}$ using grid search. During the test, the input weights with absolute values smaller than 10^{-3} were set to 0 as suggested in [124]. Moreover, for the state-of-the-art *MTFL*-based models, regularization parameter was set to be $\{10^{-4}, 10^{-3}, \ldots, 10^{3}, 10^{4}\}$. The number of features to be selected in *CMTFL-1* model was set to be $\{5, 10, \ldots, 55\}$. The number of static and dynamic features to be selected in *CMTFL-2* model were set to be $\{4, 5, 6, 7, 8\}$ and $\{5, 10, \ldots, 50\}$, respectively. Finally, the code for proposed method and the data are available at `https://github.com/picsolab/actattn`.

3.4 Results

In this section, a comprehensive set of results is presented. First, in Section 3.4.1, the forecasting effectiveness of the proposed model in comparison with the baseline and state-of-the-art forecasting approaches, and based on the aforementioned experiment settings is shown. In Section 3.4.2, different kinds of predictive features identified by the proposed model are analyzed, and their effects in relation to the theoretical factors are interpreted. In Section 3.4.3, different kinds of spatio-temporal contributions (local vs. global contributions, (inter-) local contributions) are analyzed and interpreted. Finally, in Section 3.4.4, the potential of using additional content features in the current forecasting framework is explored.

3.4.1 Performance Comparison

The forecasting performance of ActAttn was compared with the comparison methods. The results are organized to answer the following three questions:

1. Overall, how well could ActAttn forecast future protest event occurrences, compared with the baseline methods? (Section 3.4.1.1)

2. As missing information is common in social event predicting problems, how robust could ActAttn deal with missing information, compared with the baseline methods? Does ActAttn's spatio-temporal architecture help deal with the missing or noisy information? (Section 3.4.1.2)

3. How early in time can ActAttn effectively predict future protest event occurrences? (Section 3.4.1.3)

3.4.1.1 Overall Performance

As shown in Table 3.3, the results indicate that ActAttn achieves the highest F-score and Area Under Curve (AUC) values on Charlottesville (0.400 and 0.843), Ferguson I (0.462 and 0.822) and Ferguson II (0.471 and 0.853) datasets. The F-scores for all methods are low due to the imbalanced class distribution (9%–15% protest events). For instance, all SVM variants yielded 0% F-score for Ferguson I, and SVM $[s, tem, st]$ performed 0% F-score for Charlottesville. In addition to the imbalanced class distribution, different behaviors of the protest occurrences in the training and the test sets for Charlottesville and Ferguson I might be another reason for the corresponding results of the SVM variants. As a result, while

Table 3.3: Forecasting performance results.

	Charlottesville		Ferguson I		Ferguson II	
	F-score	AUC	F-score	AUC	F-score	AUC
$LR\ [tem]$	0.200	0.696	0.103	0.733	0.343	0.752
$LR\ [s, tem]$	0.182	0.789	0.259	0.766	0.327	0.789
$LR\ [s, tem, st]$	0.200	0.734	0.230	0.722	0.314	0.773
$SVM\ [tem]$	0.200	0.818	0.000	0.791	0.400	0.816
$SVM\ [s, tem]$	0.186	0.809	0.000	0.796	0.408	0.837
$SVM\ [s, tem, st]$	0.000	0.782	0.000	0.754	0.313	0.780
$LSTM$	0.240	0.752	0.415	0.801	0.417	0.819
$S + LSTM$	0.267	0.778	0.423	0.804	0.439	0.838
$S + LSTM\ (GL)$	0.308	0.793	0.423	0.805	0.440	0.839
$RMTFL$	0.182	0.663	0.250	0.703	0.250	0.829
$CMTFL{-}1$	0.182	0.664	0.350	0.711	0.316	0.805
$CMTFL{-}2$	0.200	0.661	0.333	0.711	0.324	0.815
$ActAttn\ (w/o\ GL)$	0.308	0.830	0.459	0.820	0.464	0.849
$ActAttn\ (w/o\ stAttn)$	0.324	0.797	0.406	0.783	0.409	0.842
$ActAttn\ (w/o\ spAttn)$	0.333	0.836	0.448	0.812	0.448	0.846
$ActAttn$	**0.400**	**0.843**	**0.462**	**0.822**	**0.471**	**0.853**

the protest occurrence pattern is different for each dataset (Figure 3.2), ActAttn is robust to distribution of the data and models temporal and spatial dimensions successfully.

The significance of the static features is indicated by comparing the results of $LR[tem]$ with $LR[s, tem]$, $SVM[tem]$ with $SVM[s, tem]$, and $LSTM$ with $S + LSTM$. It can be seen that, nearly in all cases, combining static features with intra-region features yields better F-score and AUC values. When the inter-region features are further combined, it is observed that $LR[s, tem, st]$ and $SVM[s, tem, st]$ give worse results compared to $LR[s, tem]$ and $SVM[s, tem]$, respectively. Thus, these models fail to capture the spatio-temporal information from the concatenated inter-region features. In our approach, combining the inter-region features with the static features and the intra-region features (i.e. combining local and global contributions) increases the performance in *ActAttn* based methods except *ActAttn (w/o stAttn)*. Moreover, $S+LSTM\ (GL)$ performs slightly better than $S+LSTM$ and eliminates some of the redundant inputs in all of the three models.

To compare the performance of ActAttn with the state-of-the-art spatio-temporal event forecasting approaches, additional experiments were performed on all the datasets with *RMTFL*, *CMTFL-1* and *CMTFL-2* proposed by [8] by employing various parameter combinations. The best test performances of these approaches on each dataset are reported. The results indicate that ActAttn significantly outperforms all three approaches on all datasets in terms of both F-score and AUC values[8].

[8] The AUC of the best model (> 0.82) suggests it is possible to rank-order or filter the states where protest events are likely to happen with reasonable accuracy.

To examine the effect of Group Lasso regularization and hierarchical attention mechanism, the performance of *ActAttn* were compared with its three variants. Although *ActAttn* slightly outperforms *ActAttn (w/o GL)*, Group Lasso regularization provides sparsity and selection of compact set of features. *ActAttn* model provides 95.0%, 76.6% and 96.8% sparsity for Charlottesville, Ferguson I and Ferguson II, respectively. It is computed as the ratio of zero input weights over total number of input connections. Furthermore, *ActAttn* was compared with *ActAttn (w/o stAttn)* and *ActAttn (w/o spAttn)* to examine the effect of hierarchical attention mechanism. It is observed that *ActAttn* significantly performs better than *ActAttn (w/o stAttn)*. This shows the importance of spatiotemporal attention layer which adjusts the local and global feature contributions. Similarly, *ActAttn* performs superior than *ActAttn (w/o spAttn)*. Removal of the spatial attention layer from the proposed architecture also results in

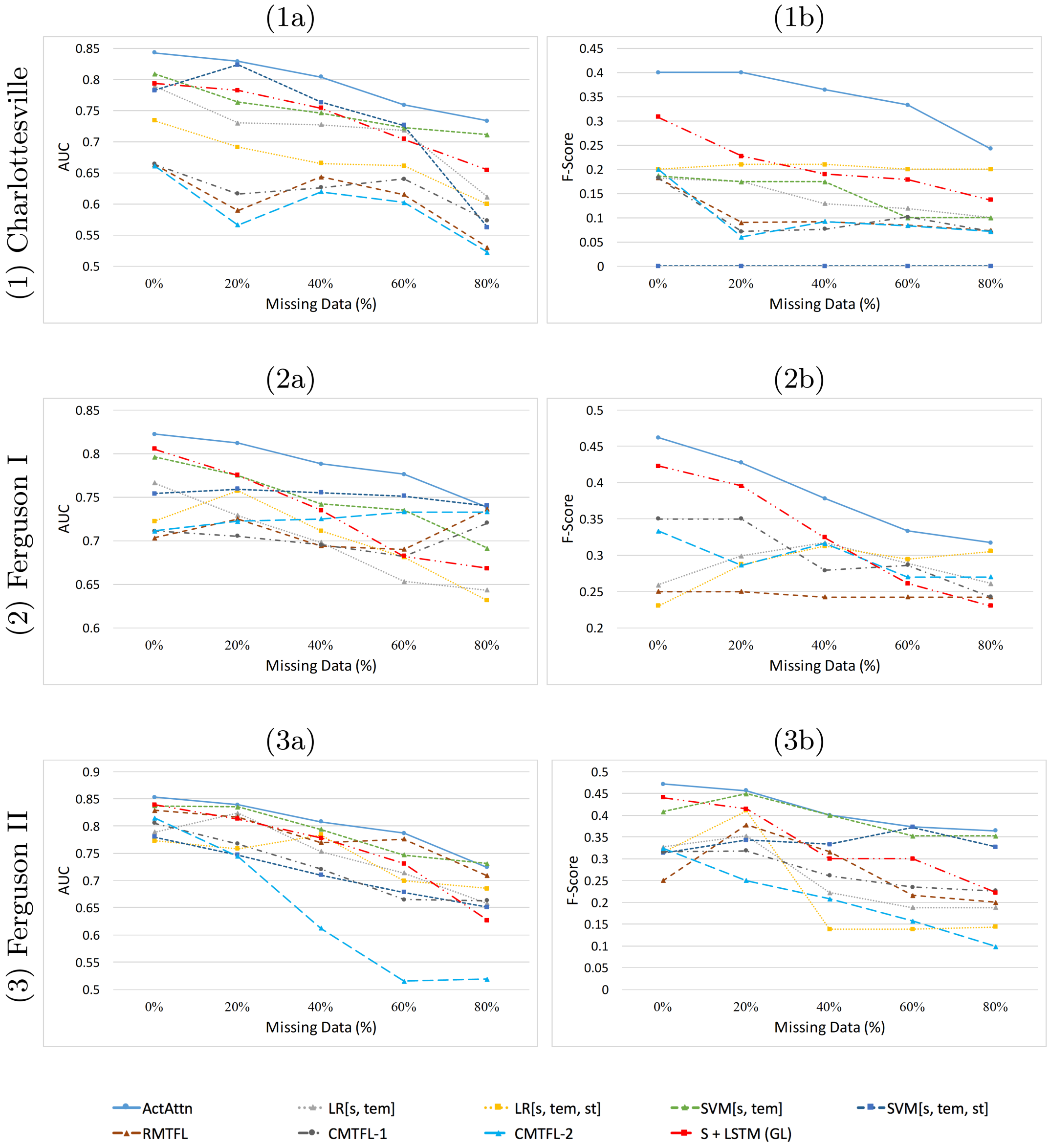


Figure 3.3: Forecasting results against varying levels of missingness (in time and space) from the test sets. The x-axes indicate the levels of missingness, and the y-axes indicate the performance in terms of (a) AUC and (b) F-Score results.

loss of interpretation capability about the most contributing locations.

3.4.1.2 Robustness to Missing Information

A common challenge in predicting/forecasting societal events is that data (including but not limited to social media data) often involve missing information or are only partially complete. For example, social media user activity may be sparse in a certain region or a particular time. As ActAttn is designed to capture the spatio-temporal characteristics and features, it is expected that ActAttn would be more robust to missing data if the model effectively captures the spatio-temporal structure from the training data. To test this, two kinds of missing information scenarios are simulated as follows:

(1) Missingness in time and space: A missing value could occur in any feature of any region at any time. To simulate this, different levels of input data (20%, 40%, 60% and 80%) are randomly removed from the test sets. An incremental approach is followed for the removal of the input data. While the level of missingness increases, the missing input data from the previous level is kept the same and additional random removal is performed for the current level. Then, the missing values are filled by randomly assigning values taking from the range of non-missing values of the corresponding features. In this setting, the comparison methods include those methods that take all features (static, temporal and spatial features) as input and have the best overall performance within each of the method variants. Figure 3.3 shows the forecasting performances of the methods for each dataset over different levels of missing data. The results indicate that ActAttn performs significantly better (in terms of both AUC and F-Score) than all the other methods on all datasets and for almost all levels of missing data.

(2) Missingness in certain regions: The missing values could occur in a particular region for an entire (short- or long-term) period of time. To simulate this, different proportions of regions (states) (ranging from 20% to 80%) are randomly selected and their input is entirely removed from the test sets. An incremental approach is followed for the removal of the regions. While the level of missingness increases, the missing regions from the previous level are kept the same and additional random removal is performed for the current level. The removed regions thus do not have any contribution to forecasting events in any of the target regions. In this setting, the methods taking features from the other locations are included for the comparison. Note that, although these methods include features from the other states, they do not differentiate intra- and inter-region (i.e. local vs. global) contributions. Therefore, it is expected that these comparison methods may suffer from missing some regional input. Figure 3.4 shows the forecasting performances of the methods for each dataset over different levels of missing region information. The results show that ActAttn outperforms the other methods in terms of both AUC and F-Score on all three datasets and for all levels of missing region information. Also, we observed that ActAttn performs more stable in nearly all conditions.

In both scenarios, it is observed that ActAttn is more robust compared to other methods. This suggests that the design of ActAttn is particularly useful in dealing with missing information – the hierarchical attention mechanism learns important regions and summarizes the spatio-temporal information from intra-region and inter-region features (local vs. global contributions), and the Group Lasso regularization imposes sparsity and selects an informative set of features. Note that the experiments for each scenario are performed for a single random

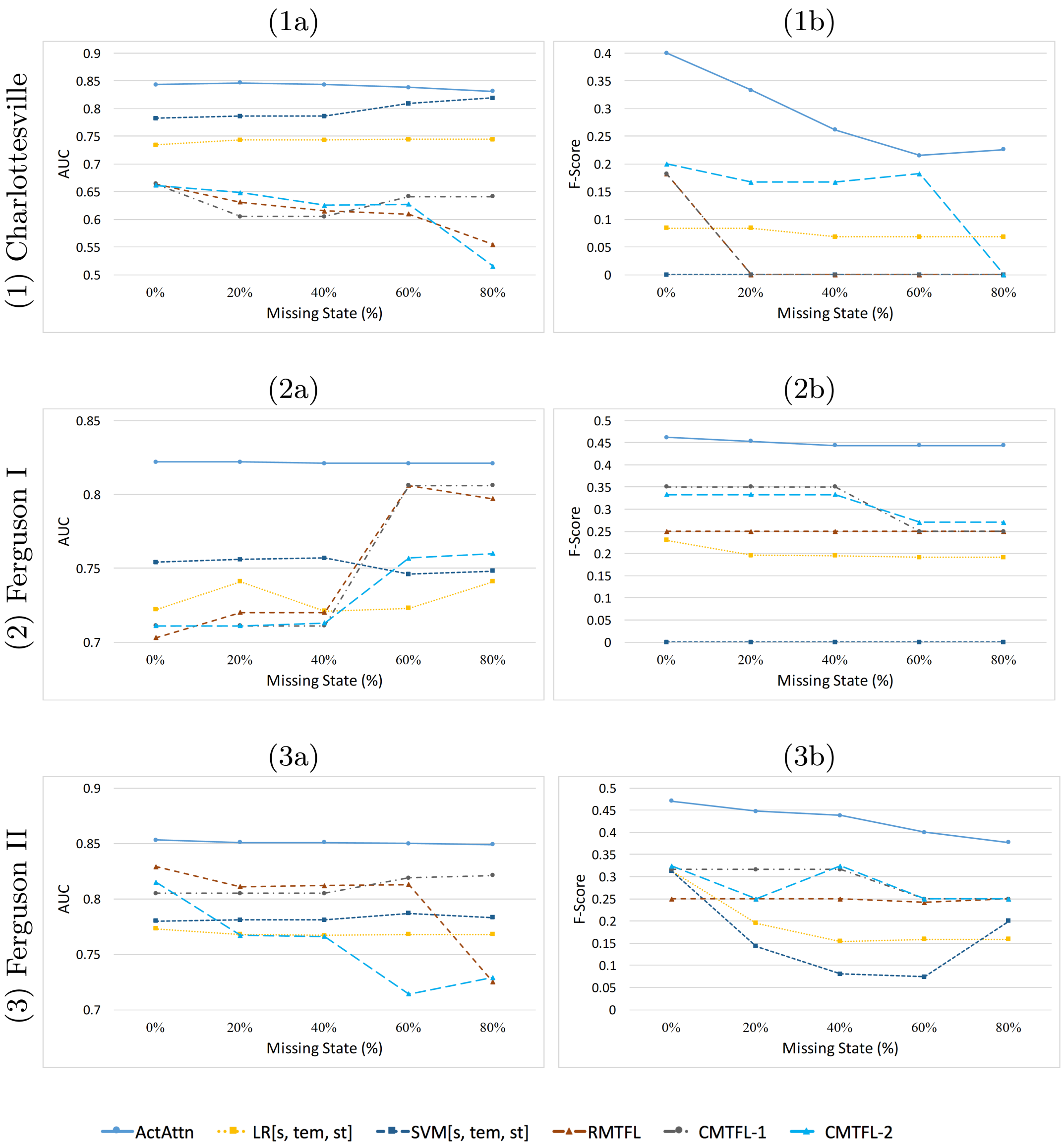


Figure 3.4: Forecasting results against varying levels of missingness for regions (states) from the test sets. The x-axes indicate the levels of missingness, and the y-axes indicate the performance in terms of (a) AUC and (b) F-Score results.

configuration. Multiple repetitions of the same experiments for both scenarios would yield more reliable results.

3.4.1.3 Performance Analysis with Varying Lead Time

To examine how early in time ActAttn effectively forecasts future protest event occurrences, the forecasting performance under different *lead time* conditions is tested. Recall that the lead time τ is the length of time (number of days in the experiments) from which the data is available for forecasting events occurring at $t + \tau$ (as defined in Section 3.1). The proposed method is evaluated with different lead time settings, where $\tau \in \{1, 2, 3\}$. Figure 3.5 shows

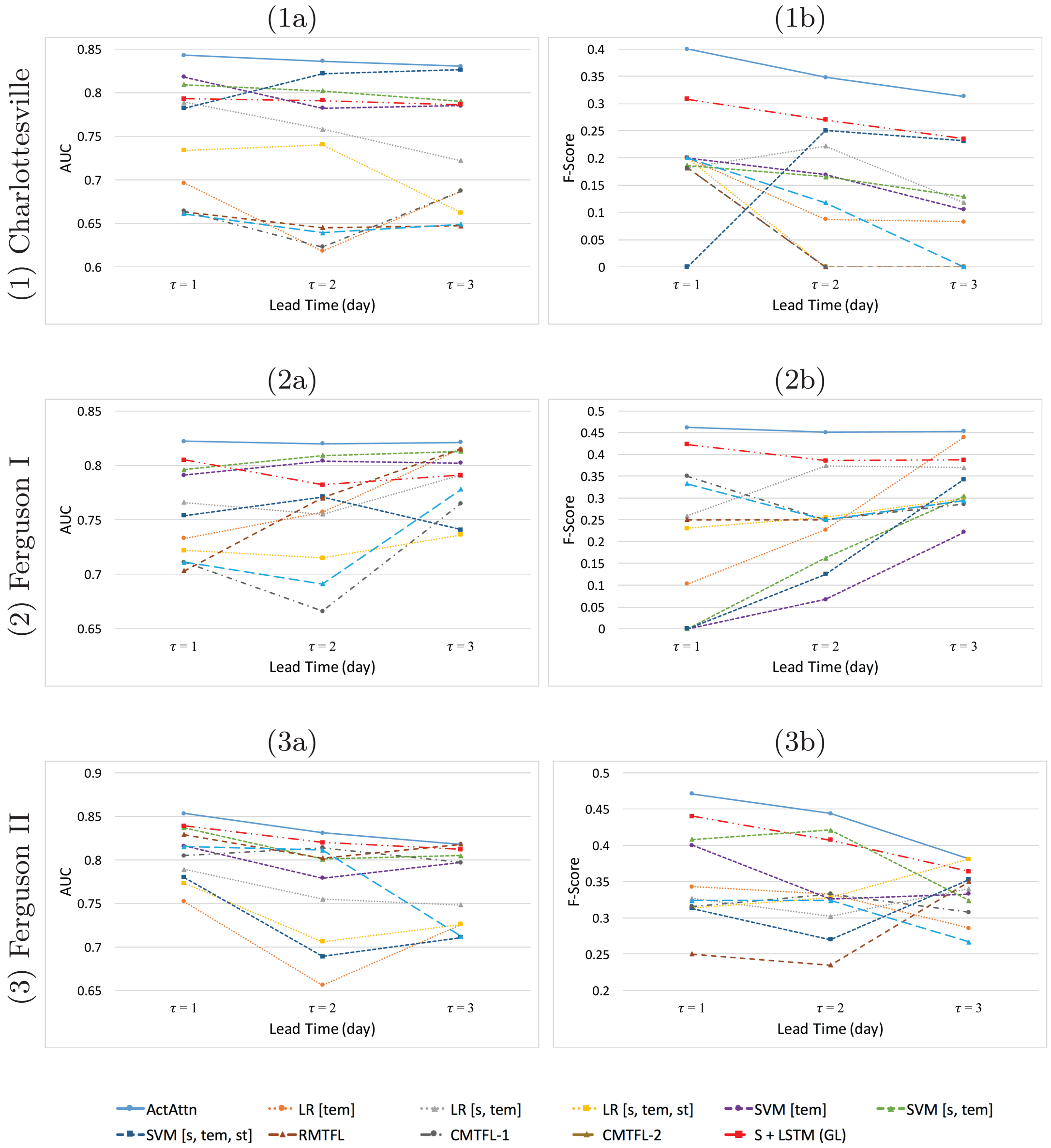


Figure 3.5: Forecasting results against different lead times. The x-axes indicate the lead time τ, and the y-axes indicate the performance in terms of (a) AUC and (b) F-Score results.

the forecasting performances of ActAttn and comparison methods over different lead time settings. The results indicate that ActAttn has significantly better performance compared to other methods in terms of AUC and F-Score on three datasets across almost all lead time settings. This suggests that ActAttn is able to achieve better and more stable performance for short-term event forecasting up to $\tau = 3$. Due to the limitation of the data, longer-term event forecasting performance is not examined in this work.

The forecasting performance of ActAttn is further examined with different window size k and lead time τ. As defined in Section 3.1, the window size represents the amount of information needed for forecasting in terms of the number of consecutive days as input. The AUC values for corresponding results are given in Table 3.4. Accordingly, the best performances are

Table 3.4: AUC results of ActAttn with respect to different window size k and lead time τ.

	Charlottesville			Ferguson I			Ferguson II		
	k = 1	k = 2	k = 3	k = 1	k = 2	k = 3	k = 1	k = 2	k = 3
τ = 1	0.842	**0.843**	0.823	0.807	0.815	**0.822**	**0.853**	0.832	0.800
τ = 2	0.839	0.836	0.823	0.807	0.820	0.820	0.831	0.836	0.832
τ = 3	0.830	0.830	0.819	0.791	0.808	0.821	0.818	0.820	0.811

achieved when ($k = 2$, $\tau = 1$), ($k = 3$, $\tau = 1$) and ($k = 1$, $\tau = 1$) for Charlottesville, Ferguson I and Ferguson II models, respectively. In general, the performance either remains stable or slightly decreases with an increase in the lead time τ regardless of window size k. It is also observed that social movements with different characteristics may require different window size length to obtain the best forecasting performance.

3.4.2 Interpreting the Impact of Features

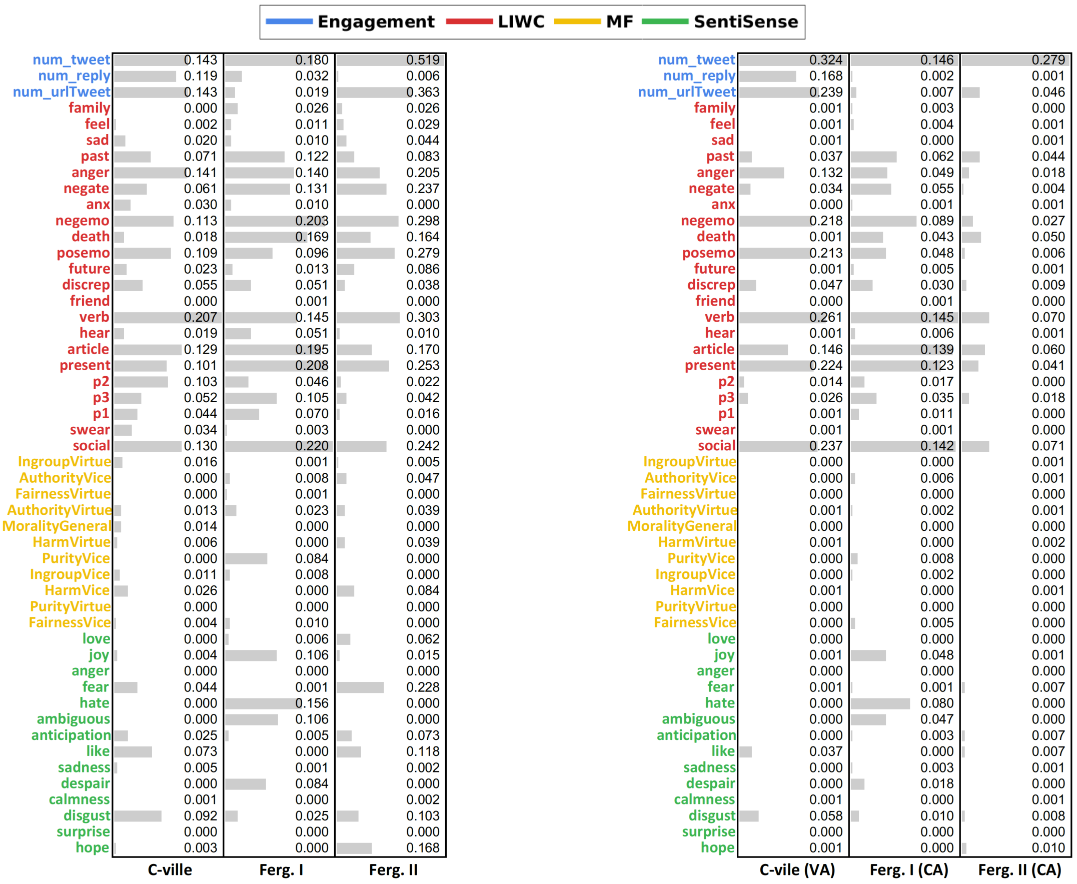
The significance of the features are interpreted by analyzing *intra-region*, *inter-region*, and *static* features. Group Lasso regularization has selected a subset of features with the most discriminative power in the models.

3.4.2.1 Intra-region dynamic features

This section analysis the importance of the intra-region dynamic features on forecasting the local future protests. In other words, the information about the input weights of (intra-) local component are provided $\mathcal{M}^{loc}$ for all three models in Figure 3.6-a. In general, most of the eliminated features in all models belong to the categories of Moral-laden and SentiSense, yet the remaining features in these categories are still important for the prediction. The detailed interpretation with respect to relationship between the key factors and the features is presented below. Also, to better understand the significance of those features in each protest contexts, a manual inspection of the tweet content was conducted.

1. Emotion. Both *positive emotion* and *negative emotion* (*posemo* and *negemo* from LIWC), are important in all models. Particularly, *anger* (from LIWC) is predictive for all, which suggests that anger is a good indicator in predicting protest for all cases. Moreover, certain particular emotions stand out for each protest scenario. For example, *disgust* (from SentiSense) is predictive in Charllottesville; *hate* (from SentiSense) in Ferguson I; and *fear* (from SentiSense) in Ferguson II.

In addition, a Moral-Laden feature, *PurityVice* (in terms of the extent of impurity and corruption) unexpectedly captures an intensely *annoying* emotion in predicting Ferguson I protests. This was uncovered when analyzing the relevant tweets in which the online community extensively express their being "*sick of*" or feeling "*disgust*" for the fact that another Black life was killed by the police.



(a) Intra-region

(b) Inter-region

Figure 3.6: Mean absolute values of intra-region and inter-region input (gate) weights. These are the input weights learned from the neural network model (the LSTM networks in the (intra-) local component and (inter-) local components), and the magnitude of weights (can take any values) allows for a comparison for the relative importance of different features. (a) intra-region input weights ((intra-) local component). (b) inter-region input weights (inter-) local components for the most contributing states.

2. Grievance. The results indicate that Moral-Laden features are not able to capture grievance. However, through further analysis of the feature *negation* (from LIWC) (the use of words such as *no*, *not*, *never*) suggests that it may serve as the indicator of grievance. This feature is important for all models, especially in Ferguson I and II. Negation is used in online community to emphasize their appraisals of how unbelievable and unrealistic when they learn about the happenings (e.g. shooting of unarmed Michael Brown, grand jury's decision on not indicting Officer Wilson, and a public rally against racism) that strongly conflict with their normal sense of moral principles. It indicates grievance referring to the feeling of illegitimate injustice.

3. Identity. *Social* (from LIWC) refers to the use of personal pronouns especially plural ones such as *we*, *you*, *they*, *people*. It is predictive for all models. These terms are extensively used to call attention on in-group members (we) to recognize the grievance and express protesting voices to out-group members (they; e.g., the police, a group considered by a majority of the online community as an embodiment of racism). It is reasonable for this feature to be significant in predicting protests and rallies, since people naturally form groups, and use those words to call for identity in such circumstances.

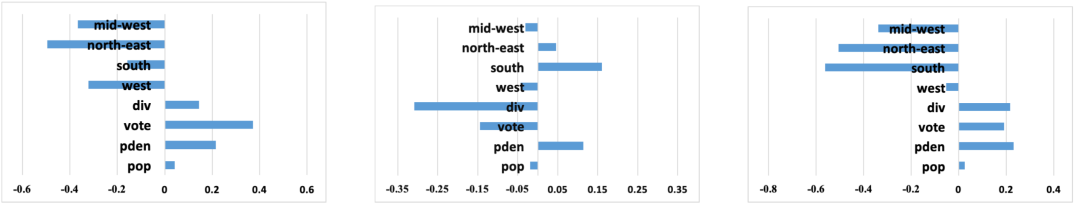
4. Social Embeddedness. Among the three relevant features (number of tweets, number of replies, and number of tweets with URLs), *num_tweets* is the most powerful for all of the three protest events. Online activism within a state is predictive of future offline protests in the same given state. *Num_urlTweet*, which indicates the amount of Twitter posts that contains an external link to other sources, is found to be a useful predictor too except for Ferguson I. This may be caused by the fact that Michael Brown's death was at first not paid much attention to among news outlets, so the external news or relevant URLs may be less indicative of online activist engagement.

5. Others. The impact of other additional features are also analyzed. The features of both *verb* (from LIWC) and *present* (from LIWC) are important in all cases. It indicates the use of verbs especially present tense of both auxiliary verbs such as *is*, *are*, *have*, *can* to emphasize the happenings and perceived grievance as a serious matter of fact, as well as the use of action verbs such as *go*, *take*, *make*, *need*, *think* that call for necessary actions.

The features of *personal pronouns* (from LIWC) are significant predictors as well. They involve the reference of and the discussion on certain person/people at the center of why people protest for or against. For example, *you* is important for Charlottesville; the second-person pronoun extensively refers to President Trump as online activists questioned him earnestly about his position in racism. *he* is important in predicting Ferguson I protests, which is used to refer mostly either Michael Brown or Eric Garner, both of whom were killed by the police. *they* mostly refers to the police. In Ferguson II, online activists focused more on the judicial system who had been believed unsuccessful in bringing about justice, thus personal pronouns are less predictive.

3.4.2.2 Inter-region dynamic features

The effectiveness of the inter-region dynamic features is explored by analyzing the input weights (only the portions which connect inputs to input gates) of each (inter-) local component in the global component, $\mathcal{M}^{glob}$. Figure 3.6-b summarizes the importance of the inter-region dynamic features in predicting protest within given states. Large percentages (96.5%, 77.6%, and 97.9% in cases of Charlottesville, Ferguson I and Ferguson II, respectively) of the input weights are discarded as a result of Group Lasso regularization. Virginia (VA) from Charlottesville model, California (CA) from Ferguson I model and CA from Ferguson II model are selected models to analyze the inter-region input weights because these states are all 'hub' states (with the largest contribution to prediction than the others') for the corresponding models (explained in Section 3.4.3). In general, the result suggests that Group Lasso selects a set of features which are similar to the ones of intra-region features for all models although the number of selected inter-region features are less than intra-region features. This indicates that the feature contributions from the local and the global exhibit similar patterns. *num_tweet*, *past*, *anger*, *negemo*, *verb*, *article* and *social* from LIWC are the common important inter-region features for all models. Also, for all models, *num_tweet* is the most important inter-region feature, which indicates online community activities in other states could be also significant across all other states.



(a) Charlottesville Model (b) Ferguson I Model (c) Ferguson II Model

Figure 3.7: The values of the static feature weights. These are the static feature weights learned from the neural network model. The weights (can take any values) allows for a comparison for the relative importance of different features. (a) Charlottesville model. (b) Ferguson I model. (c) Ferguson II model.

3.4.2.3 Static features

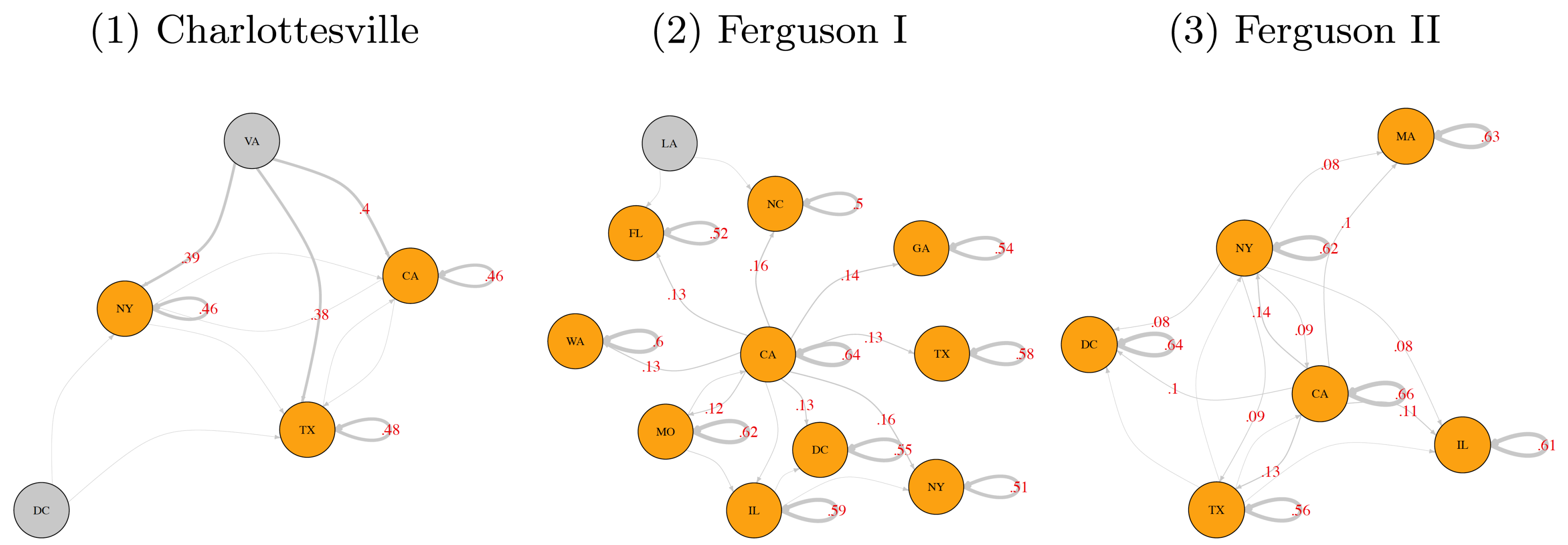
Figure 3.7 shows the importance of static feature weights in the three models. These feature weights contribute to prediction in the final layer of the neural network. The static features representing the U.S. regions indicate how predictive the coarser grained region class for a given state is (e.g., is a state in the South more likely or less likely to have future protests?). The results of Charlottesville and Ferguson II models exhibit similar patterns, which suggests that protest events in both social movements took place more all over the U.S., while Ferguson I started locally with a majority of Black communities, and its model shows that being a Southern state itself is predictive for possible future protests. For the Charlottesville and the Ferguson II models, being in the regions *mid-west* and *north-east* of the U.S., population density and vote behavior are important indicators for a state to forecast future local protests.

3.4.3 Interpreting the Local and Global Contributions and Hubs

ActAttn enables us to explore the proportion of the local (intra-region) and the global (inter-region) contributions in forecasting protest events, and allows for discovering the "hubs" that have a more salient contribution in predicting protest events globally. The intra-region and the inter-region contributions can be identified based on the spatiotemporal attention weights in the model, and the hubs can be identified as the regions (states) whose inter-region contributions to others are significant. In this study, it is observed that the spatial attention weights do not differ significantly across the different samples since the spatial attention layer learns the locations with higher contributions to forecasting regardless of the target location. These weights represent an overall, consistent spatial relationship among the regions and across the days. Therefore, in the following analyses, the results aggregated from both all the test samples and the representative test samples are presented.

3.4.3.1 Local vs. global contributions

To examine the differences between the local (intra-region) and the global (inter-region) contributions for forecasting events, a contribution graph for each model is created. As shown in Figure 3.8, the orange nodes represent states where the offline events are correctly predicted



(1) Charlottesville

(2) Ferguson I

(3) Ferguson II

Figure 3.8: Exploration of local and global contributions to forecasting. While the orange nodes represent the states which are correctly predicted by the corresponding models, the gray nodes denote the states either not correctly predicted or where no events occurred, yet still contribute to forecasting events in the correctly predicted states with a value above a certain threshold. The edges indicate the contribution to forecasting from source state to target state. The thicker edge, the more contribution.

by the model, the gray nodes represent the states where either the events are not correctly predicted or no event occurred, yet still contribute to forecasting events in other states. For visual clarity, only gray nodes having an inter-region contribution greater than a certain threshold (0.01, 0.05 and 0.01 for Charlottesville, Ferguson I and Ferguson II, respectively) to any of the orange nodes are indicated. An edge indicates the contribution of forecasting for a target state from a source state and the edge weight (thickness) reflects the contribution magnitude. Also for visual clarity, only edges whose weights are more than a certain threshold, which is 0.05, 0.1 and 0.05 for Charlottesville, Ferguson I and Ferguson II, respectively, are shown. For a target state, the self-loop represents the intra-region contribution where the other incoming edges represent the inter-region contributions to that state. Note that there might be states where the protest events occurred on multiple test days. For such states, the average contributions of all these days are taken for summarization, and they are shown in the graph.

The hierarchical attention mechanism in the proposed ActAttn model enables a systematic way to interpret the intra-region (local) and inter-region (global) contributions. The contribution from a source state to a target state (inter-region) at a specific event day is calculated by $(\beta_{glob} * \alpha_{source})$ where β_{glob} is the attention weight corresponding to the global component $\mathcal{M}^{glob}$, and α_{source} is the spatial attention weight for the source state in the global component, $\mathcal{M}^{glob}$. Similarly, the intra-region contribution can be estimated by $(\beta_{loc} + \beta_{glob} * \alpha_{target})$ where β_{loc} is the attention weight corresponding to the (Intra-) local component and α_{target} is the spatial attention weight for the target state in the global component. As shown in Figure 3.8-(1), VA has a salient contribution (as a part of the global contribution) to forecast the states where the events are correctly predicted for Charlottesville case. In other words, social media activity in VA, would be a powerful signal for forecasting the offline events in the other states. Moreover, CA (mostly), IL and MO can be regarded as hubs as they contribute more than others to the target states for forecasting events in Ferguson I (Figure 3.8-(2)). On the other hand, the inter-region contributions from CA and NY to target states are much more than the other states in Ferguson II (Figure 3.8-(3)). Note that the local (intra-region) contributions (reflected by the self-loop weights) for any target state are higher than the contributions from any other state in all three models. This suggests that local activity still plays a more important role than the activity of any other states. Interestingly, in the case of Charlottesville, the global contribution (the total inter-region contributions of all other states) for a target state is more than the local one, suggesting that the Charlottesville protests have a very distinct spatio-temporal process compared with the other two cases.

3.4.3.2 The effect of hubs

To further illustrate the hub effect, the representative test samples are selected for Texas (TX), Washington (WA) and Illinois (IL), which are correctly predicted events by Charlottesville, Ferguson I and Ferguson II models, respectively.

In the Charlottesville model, the spatiotemporal attention weights for the local and the global contributions are 0.458 and 0.542, respectively. It means that the global component $\mathcal{M}^{glob}$ contributes more to forecasting the protest in TX for the given sample. To further analyze the global contribution and the hub effect, the inter-region input (gate) weights (input weights in (inter-) local components) and the spatial attention weights are visualized as given in Figure 3.9. It is observed that Group Lasso regularization selects the informative features from

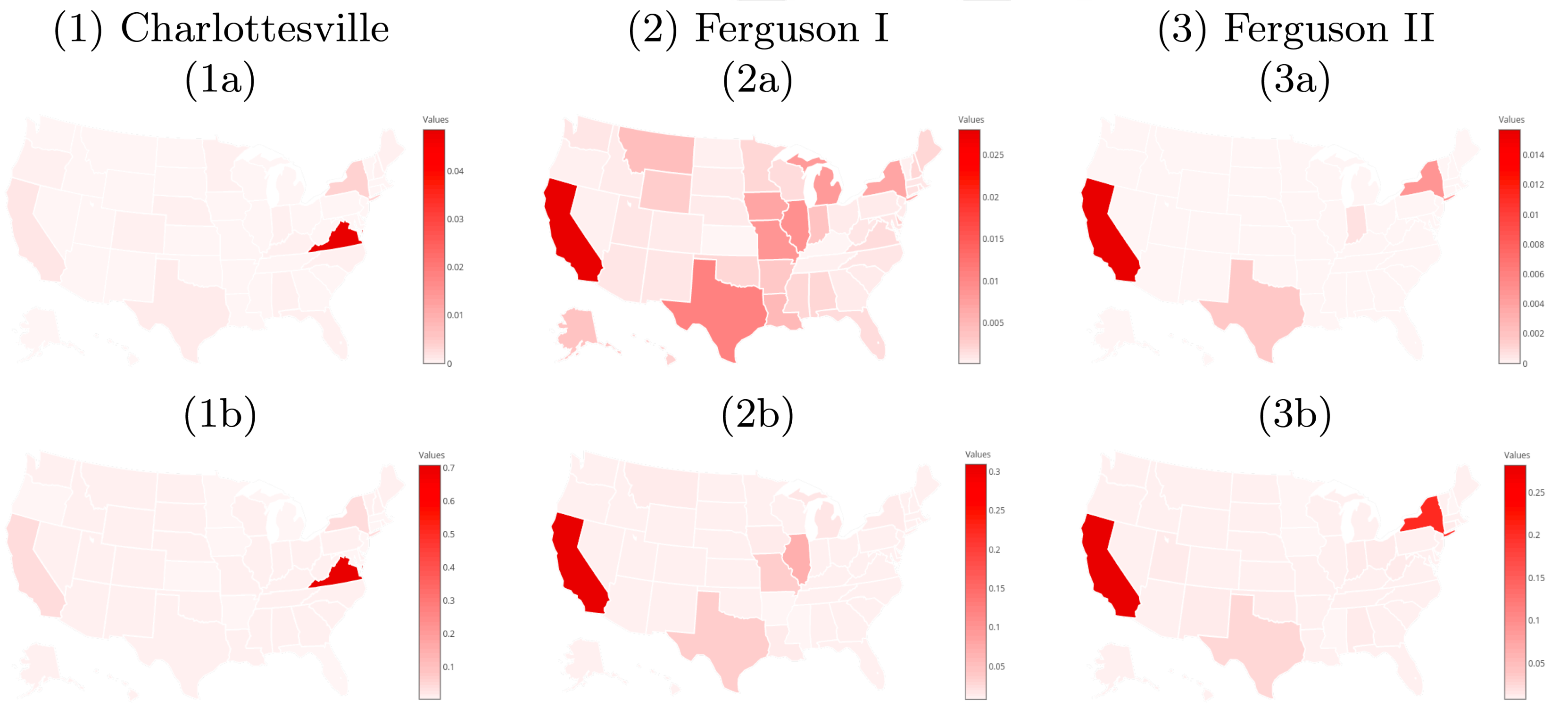


Figure 3.9: Exploration of global contribution and hub effect. (a) Mean absolute values of the inter-region input weights across the states. (b) Attention weights of the spatial attention layer for predicting the protests in TX (1b), WA (2b), and IL (3b).

only a few states, namely VA, New York (NY), CA and TX (Figure 3.9-1a), and the spatial attention layer further selects VA, CA and NY as “hubs” (Figure 3.9-1b). VA is the most contributing ‘hub’ in predicting protest event for the given test sample from TX. Since the trigger event of Charlottesville Rally occurred in VA, higher attention weight for VA is the potential indicator that the proposed model is able to model spatio-temporal relationships among the regions successfully for the Charlottesville dataset.

In the Ferguson I model, the spatiotemporal attention weights for the local and the global contributions are 0.591 and 0.409, respectively. This indicates that locality is more predictive for the given test sample of WA. Moreover, the spatial attention attends the states CA, IL, Missouri (MO) and TX (Figure 3.9-2b), suggesting the high impact of these states. Ferguson is located in St. Louis, MO where the shooting of Michael Brown happened, which is the starting point of the protests. It is also very close to IL border. The reactions to Ferguson shooting in the social media is most likely started to spread from these states. CA is an active state where both online (tweet volume) and offline activities (protests) occurred much more frequently than the other places. Therefore, these could be the explanations why these states are hubs in this model.

In the Ferguson II model, in predicting the protests in IL, the spatiotemporal attention weights for the local and the global contributions are 0.576 and 0.424, respectively for the correctly predicted test sample from IL. As shown in Figure 3.9-3a and Figure 3.9-3b, CA and NY are selected more by the spatial attention as the most attended regions (among those initially given by the Group Lasso). This suggests that the protest forecasting may be impacted by the heightened social media discussion in these hub states, in relation to, e.g., the NYPD shooting of Akai Gurley, and the arrest of BLM activists in the Bay Area during the study period.

3.4.4 Testing Predictive Power with Additional Features

While the selection of features is theory-driven, the possibility of incorporating additional features (which are emerging from the events unfolding) should also be considered since they could help increase the predictive power of the proposed model in a meaningful way. For example, specifically, it is considered whether there are keywords used by the Twitter users to plan, organize, or mobilize protests that may also serve as effective features. Since in most cases mobilization activities and activism on Twitter are organized and advocated by the Twitter users through hashtags, the mostly used hashtags are taken into consideration. For this purpose the top-k ($k = 100$) hashtags based on TF-IDF values are analyzed. Each day is treated as a separate document. Then, these top-100 keywords are included as additional features to see if they affect on forecasting, and analyzed if there are predictive ones.

The ratio of number of tweets that include the hashtag to the total number of tweets at the specific time (day) is assigned as the feature value for the corresponding hashtag. According to the results given in Table 3.5, employing the additional features decreases the performances in terms of both F-score and AUC for all three datasets. Furthermore, the importance of these hashtag features are explored by analyzing the input weights. In all of the three cases, less than 10% of the features have non-zero weights after Group Lasso regularization, which means that most of the features do not have any contribution to forecast events as both intra-region and inter-region features. The informative hashtags include “#theresistance" for Charlottesville, “#ferguson", “#mikebrown" and “#justuceformikebrown" for Ferguson I, and “#ferguson",

Table 3.5: Forecasting performance results without and with hashtag features. C.F. stands for content features.

	Charlottesville		Ferguson I		Ferguson II	
	F-Score	AUC	F-Score	AUC	F-Score	AUC
Without C.F.	**0.400**	**0.843**	**0.462**	**0.822**	**0.471**	**0.853**
With C.F.	0.308	0.814	0.453	0.815	0.435	0.825

“#ericgarner", “#tamirrice" and “#fergusondecision" for Ferguson II. However, the weights of these features are much less than the weights of those theory-driven features that are first employed in the original model.

3.5 Discussion and Conclusion

In this chapter, an interpretable, spatio-temporal predictive model, called ActAttn, was presented to forecast future societal events. The proposed model was applied to forecast offline protest events from online activities as one of the application domains for societal events. A novel deep learning architecture was developed, which models the local and the global dynamic information concurrently as well as the location-specific static information through a hierarchical attention mechanism. ActAttn also enables interpretation in both local and global dimensions, and theory-relevant activity features. Through extensive experiments, the strength of the proposed model was demonstrated. The proposed model was compared with the baseline and state-of-the-art methods, and it achieved a superior forecasting performance for all three movement datasets. It has also been more robust to missing data and consistently outperformed other methods in various early forecasting settings.

In the proposed architecture, Group Lasso regularization and hierarchical attention mechanism were employed to perform theory-driven feature selection, and explore and identify the most contributing locations to forecast the protest events. ActAttn both allows us to examine the important intra-region and inter-region features (which are predictive for future protest events), and is capable of differentiating the intra-region (local) and inter-region (global) contributions. In general, for all social movement datasets it was observed that, LIWC-related features and Twitter engagement features are the most informative feature groups in common. Moreover, a few of the locations (states) had more salient contribution to forecasting local events, and these locations differ in each model.

In relation to theory-driven features, it was observed that the greater volumes of tweeting and networking behaviors for social embeddedness (including original tweets, replies, and associating content with hyperlinks) had high predictive power. This result is consistent with prior empirical studies (e.g., [19]). The negative *emotion*s have been studied and theorized to be associated with protests [16, 14], and the experiment results are consistent with the literature, particularly for anger. However, other specific negative emotions, such as disgust, hate, and fear also stood out, and showed distinct predictive power for the Charlottesville counter-protests, Ferguson I and Ferguson II, respectively. Moreover, *grievance* was not captured through the Moral-Laden based features. However, it was discovered that negation (from the

LIWC dictionary) could be a good predictor feature for all protest cases. By manual inspection of sampled tweets, it was revealed that its semantic meaning could serve as an indicator of grievance. This could be a potential to identify information of grievance in future relevant studies. Finally, for the key factor *identity*, social category from the LIWC dictionary was able to capture the group identities, such that it is predictive for Charlottesville and Ferguson I, but not Ferguson II. In brief, the proposed model goes beyond indicating that online discussion, including emotional tweets, may help predict offline protests. That point has been studied and widely recognized. Rather, this study offers insights as to where (intra-region or inter-region) and how (the features were not selected randomly or through unsupervised learning, but theory-driven) the features may offer explanatory power.

In relation to hub locations (states), it was observed that only a few locations had more salient contribution to forecasting future protest events in the target locations, and they were different for each model. For Charlottesville model, VA has more contribution to prediction which can be explained by that the trigger event for of Charlottesville rally occurred in VA. Note that its contribution to forecasting future protests is more than the contribution of local dynamics in every target location. Moreover, CA, MO, IL and TX are the most contributing set of states for the prediction of future protests in Ferguson I model. The shooting of Micheal Brown happened in Ferguson (in the border of MO-IL) and the reactions have been spread across the country in terms of both online and offline activities. This explains why the proposed model selects these states as the hubs. Lastly, CA, NY and TX are the detected hubs for Ferguson II model, which can be explained by high attention in social media related to the police shooting of Akai Gurley in NY, and the arrest of BLM activists in CA during the study period.

3.5.1 Limitations

There are several limitations in the proposed work. (1) The results indicated that considering spatial relationships among the locations increases the performance of forecasting protest events. However, the proposed architecture models the spatial structure irrespective of the locations of events. In other words, it does not differentiate the pairwise relationship between a particular event location and other locations. Future research might consider modeling the relationships between pairs of locations. (2) In the context of forecasting protest or other civil unrest events, data is generally sparse in terms of event occurrences. Events either increasingly happen within a short period after a trigger event, or only occur in particular locations. The data sparsity makes it difficult to learn complex spatio-temporal relationships. The current proposed model was not specifically designed to tackle this data sparsity issue. (3) In the current architecture, the global component $\mathcal{M}^{glob}$, which models the spatio-temporal relationships over locations, is a complex component. It consists of a (inter-) local component for each location where each component has its own LSTM component. As the number of locations increases, the number of parameters to be learned increases linearly. Although Group Lasso regularization has significantly reduced the complexity of this component, to further reduce the complexity of the model would be more desirable.

CHAPTER 4

CASTNET: COMMUNITY-ATTENTIVE SPATIO-TEMPORAL NETWORKS FOR FORECASTING SOCIETAL EVENTS

In this chapter, an interpretable, community-attentive, spatio-temporal forecasting model, named CASTNet, is proposed. Assuming that different locations could share similar dynamics and inspired by the idea of multi-head attentional networks [121], the proposed method aims to learn different representation subspaces of cross-regional dynamics, where each subspace involves a set of locations called “community” that share similar behaviors. The proposed architecture is called “community-attentive" since it allows the prediction for a given location to be individually optimized by the features contributed by a mixture of communities. Specifically, combining the features of the given target location and features from the communities (referred to as local and global dynamics), the model learns to forecast the number of societal events in the target location. The proposed method differentiates the global contributions from the communities with respect to the target location. Meanwhile, by leveraging Group Lasso regularization [124] and hierarchical attention mechanism, the proposed method allows for interpreting what local and global features are more predictive, what communities contribute more to forecasting incidences at a location, and what locations contribute more to each community.

The proposed model is evaluated on the domain of opioid overdose events, where the purpose is to forecast future opioid overdoses from spatio-temporal crime dynamics. The literature has highlighted the relationships between the opioid use and crime incidents with different aspects including cause (that opioid use leads to criminal activities [32]), effect (that involvement in criminal behavior leads to drug use [31]), and common causes (that crime and drug use tend to co-occur [34]). Crime occurrences also have non-trivial spatio-temporal characteristics – for example, routine activity theory suggests that crimes may exhibit spatio-temporal lags as the *likely offenders* of one place may reach *suitable targets* in *other places*. Given the plausible relationship between the crime dynamics and opioid use as well as the availability of real-time crime data for various locations, the proposed model is validated on two real-world opioid overdose datasets through extensive experiments. Also, in-depth analysis and comparison across several baselines and state-of-the-art methods are performed.

This chapter first defines the forecasting problem on a specific societal event domain (i.e. opioid overdose forecasting), where the proposed method is applied and evaluated. Next, the architecture of the proposed model is presented in detail. Then, the experiment details are given including datasets, extracted features, comparison methods and experimental settings. After that, the experiment results are provided including in-depth analysis and comparison across several baseline and state-of-the-art methods. Finally, the discussion and conclusion about the proposed methods are presented including the limitations of the current work.

4.1 Problem Definition

Suppose there are L number of locations-of-interest (e.g. neighborhoods, districts) and each location l can be represented as a collection of its static and dynamic features. While the static features (e.g. demographics, economical indicators) remain same or change slowly over a longer period of time, the dynamic features are the updates for each time interval t (e.g. day, week). Let X_l^{stat} be the static features of location l, and $X_{t,l}^{dyn}$ the set of dynamic features for location l at time t. We are also given a continuous variable $y_{t^*,l} \in \mathbb{N}$ that indicates the number of opioid overdose incidents (e.g. emergency medical services (EMS) calls, deaths) at location l at future time t^*. The collection of dynamic features from all locations-of-interest within an observing *time window* with size w up to time t can be represented as $\mathcal{X}_{t-w+1:t}^{dyn} = \{\mathcal{X}_{t-w+1}^{dyn}, \ldots, \mathcal{X}_t^{dyn}\}$, where $\mathcal{X}_{t'}^{dyn} = \{X_{t',1}^{dyn}, \ldots, X_{t',L}^{dyn}\}$.

The purpose is to predict the number of future opioid overdose incidents $y_{t^*,l}$ at specific location l at a future time $t^* = t + \tau$, where τ is called the *lead time* for forecasting. The forecasting is based on the static and dynamic features of the target location itself, as well as the dynamic features in the environment (from all locations-of-interest). Therefore, the forecasting problem can be formulated as learning a function $f(X_d^{stat}, \mathcal{X}_{t-w+1:t}^{dyn}) \rightarrow y_{t^*,d}$ that maps the static and dynamic features to the number of opioid overdose incidents at the future time t^* at a *target* location d.

To facilitate spatio-temporal interpretation of the forecasting, it is sought to develop a model that can differentiate contribution of the features, the locality (local features vs. global features) and the importance of latent communities when contributing to the prediction of other locations. Therefore, the dynamic features $\mathcal{X}_{t-w+1:t}^{dyn}$ are further organized into two sets: the *local* features, $\{X_{t-w+1,d}^{dyn} \ \ldots, X_{t,d}^{dyn}\}$ represent dynamic features for the target location d, and the *global* features, $\{X_{t-w+1,l}^{dyn} \ \ldots, X_{t,l}^{dyn}\}$ for $l \in \{1, 2, \ldots, L\}$, contain the sequences of dynamic features for all locations of interest.

4.2 Proposed Architecture

As shown in Figure 4.1, the proposed architecture consists of three primary components, the local component (Figure 4.1-a), the static component (Figure 4.1-b) and the global component (Figure 4.1-c). The global component is designed to model the global contribution of dynamic features for all-locations-of-interest by learning different representation subspaces of global dynamics, and to output target location-specific global contribution. On the other hand, the local component is designed to model the contribution of local dynamic features for the target location. Finally, the static component models location-specific static information about the target location.

4.2.1 Global Component

This component produces the target location-specific global contribution (from all locations) to forecast the number of incidents at the target location d at future time t^*. It consists of K number of community blocks, where each community block learns a different representation subspace of global dynamic features, which is inspired by idea of multi-head attention [121].

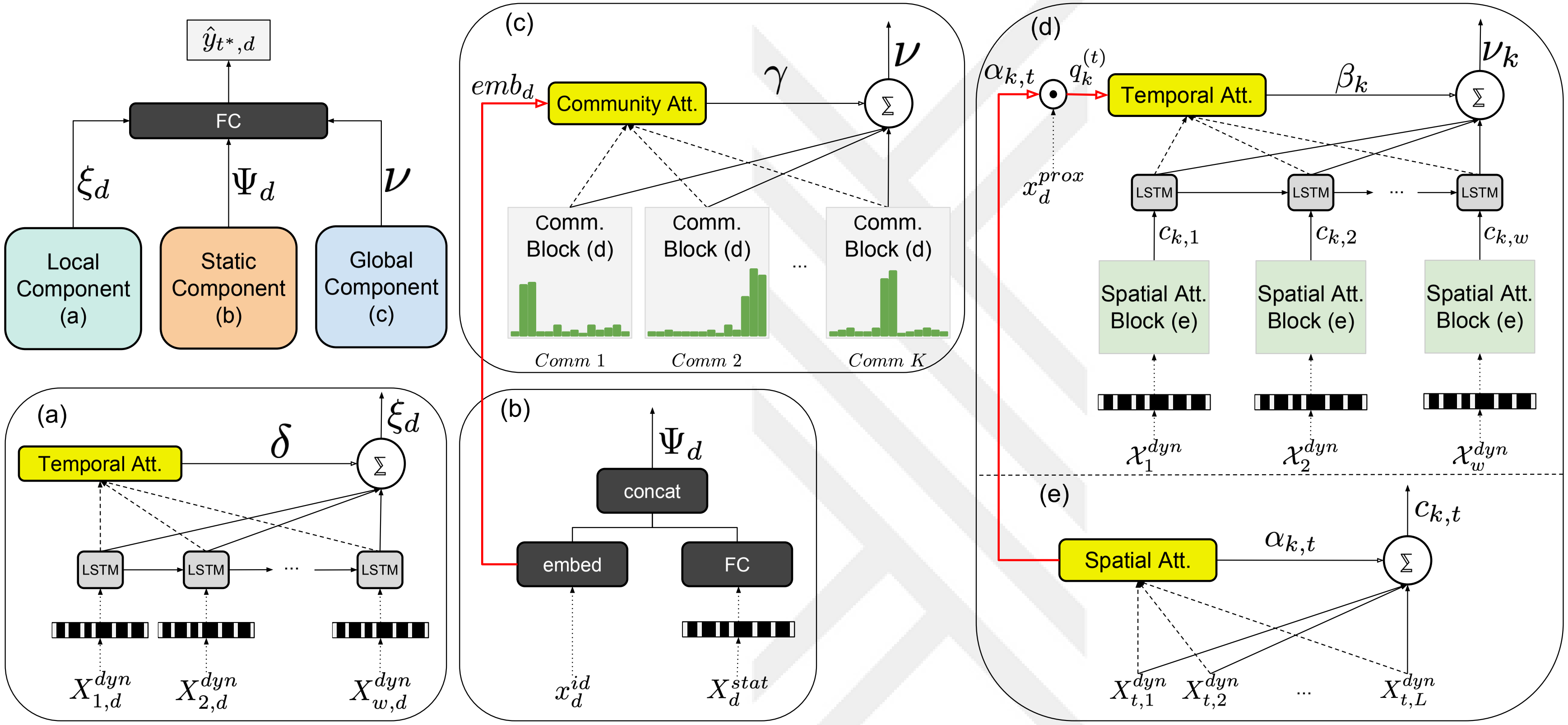


Figure 4.1: Overview of our proposed CASTNet architecture [136]. The local component (a) models local dynamics of the locations, and the static component (b) models the static features. The global component (c) summarizes different representation subspaces (i.e. communities) of global dynamics, learned by community blocks (d), by querying these multi-subspace representations through the embedding of the target location (emb_d). Spatial Att. Block (e) reweights the global dynamics of locations. Checkered rectangles on top of the inputs represent GL regularization. Red arrows indicate the queries for the corresponding attentions. “FC": fully-connected layer; “embed": embedding layer, “concat”: concatenation layer.

A community block (Figure 4.1-d) models the global dynamic features through a hierarchical attention network which consists of a spatial attention block (Figure 4.1-e), a recurrent unit and a temporal attention. For the sake of clarity, the internal mechanism of global component is explained in a bottom-up manner by following the order (Figure 4.1-e $\rightarrow$ 4.1-d $\rightarrow$ 4.1-c):

Spatial Attention Block is used to reweight the contribution of dynamic features of each location i at time t. More specifically, the attention weight, $\alpha_{k,t}^{(i)}$, represents the feature contribution of the location i at time t to the community k. With this attention block, the proposed architecture allows for forming communities, and reweighting feature contributions of the members to the corresponding communities. Since higher spatial attention weight for a location indicates the involvement of its dynamic features in this community, it is called *community memberships*. $c_{k,t}$ is the context vector, which summarizes the aggregated contribution of all locations as follows:

$$e_{k,t} = (v_k^{glob})^\intercal tanh(W_k^{glob} \mathcal{X}_t^{dyn} + b_k^{glob}), \tag{4.1}$$

$$\alpha_{k,t}^{(i)} = \frac{exp(e_{k,t}^{(i)})}{\sum_{l=1}^{L} exp(e_{k,t}^{(l)})}, \tag{4.2}$$

$$c_{k,t} = \sum_{l=1}^{L} \alpha_{k,t}^{(l)} X_{t,l}^{dyn}, \tag{4.3}$$

where $W_k^{glob} \in \mathbb{R}^{n \times n}$, $b_k^{glob} \in \mathbb{R}^n$ and $v_k^{glob} \in \mathbb{R}^n$ are the parameters to be learned, and n is the dynamic feature size of any location. $e_{k,t}$ keeps the scores for the contribution of dynamic features of the locations in the community k at time t. After the context vector $c_{k,t}$, which is the aggregated contribution of dynamics of all locations for the community k at time t, is computed, it is fed to the recurrent unit.

Recurrent unit is used to capture the temporal relationships among the reweighted global dynamic features for the community k as follows:

$$h_{k,t} = f_k(h_{k,t-1}, c_{k,t}), \tag{4.4}$$

where $f_k(.)$ is the non-linear activation function for community k, and $h_{k,t}$ is the t-th hidden state of k-th community. LSTM [111] is used in the proposed model (in each community block) to capture the temporal relationships among the dynamic features. It has been shown effective in capturing potential temporal dependency [137, 114], and it addresses the vanishing and exploding gradient problems of basic RNNs by using explicit gating mechanisms (input, output and forget gates) to regulate the memory updates.

Temporal Attention is applied on top of the LSTMs to differentiate the contribution of latent representations of global dynamic features at each time point and for each community. To make the output specific to target location, a *query* scheme is incorporated based on a time-dependent community memberships (i.e., contribution of each location to the community) where the membership is further reweighted based on the location's spatial proximity to target location (with nearby locations getting larger weights than further ones). Specifically, let $\beta_k^{(i)}$

denotes the attention weight over the hidden state $h_{k,i}$ of community k at time i. The context vector ν_k, which is the aggregated contribution from community k, can be learned through the proximity-based weighting scheme as follows:

$$q_k^{(i)} = x_d^{prox} \bullet \alpha_{k,i}, \tag{4.5}$$

$$\beta_k^{(i)} = \frac{exp(q_k^{(i)})}{\sum_{t=1}^{w} exp(q_k^{(t)})}, \tag{4.6}$$

$$\nu_k = \sum_{t=1}^{w} \beta_k^{(t)} h_{k,t}, \tag{4.7}$$

where $x_d^{prox} \in \mathbb{R}^L$ is a vector encoding the proximity of the target location d to all locations. $(\bullet)$ is the dot product operation. q_k^i is the score for the contribution at time i. Here, the proximity of two locations is calculated based on the inverse of geographic distance, particularly *haversine (hvrsn)* distance:

$$prox(l_1, l_2) = \frac{1}{\sqrt{1 + hvrsn(l_1, l_2)}}, \tag{4.8}$$

$$hvrsn(l_1, l_2) = 2R \times arcsin\left(\sqrt{sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + cos(\varphi_1)cos(\varphi_2)sin^2\left(\frac{\varrho_2 - \varrho_1}{2}\right)}\right), \tag{4.9}$$

where R is the radius of the Earth, (φ_1, ϱ_1) and (φ_2, ϱ_2) are the latitude and the longitude pairs for the locations l_1 and l_2, respectively. Note that the inverse of the haversine distance is used for the proximity definition between two locations in this work. However, any kind of proximity depending on the application domain can be defined and integrated into the proposed method easily.

Community Attention generates the output of the global component. It aims to produce a global contribution with respect to the target location d by combining different representation subspaces for each of the communities $\{\nu_1, \nu_2, \ldots, \nu_K\}$. In other words, the contributions from the different subspace representations of the global dynamic features are reweighted based on the target location d. A soft-attention approach is then employed to combine the contributions from all K communities. Here, to make the prediction specific to the target location, we incorporate a *query* scheme, which takes each community vector ν_k as *a key* and the embedding of the target location emb_d as a *query*, as follows:

$$u_k = r^{\intercal} tanh(V\nu_k + emb_d), \tag{4.10}$$

$$\gamma^{(i)} = \frac{exp(u_i)}{\sum_{k=1}^{K} exp(u_k)}, \tag{4.11}$$

$$\nu = \sum_{k=1}^{K} \gamma^{(k)} \nu_k, \tag{4.12}$$

where u_k is the score for the output of the community k, $V \in \mathbb{R}^{m \times m}$, and $r \in \mathbb{R}^m$ are the parameters to be learned, m is the number of hidden units in LSTMs in the community blocks, and ν is the final output of the global component with respect to the target location d.

4.2.2 Local Component

This component is designed to model the contribution of the local dynamic features for any target location d (Figure 4.1-a). It basically includes a recurrent unit and a temporal attention that focuses on the most informative time instants. The dynamic features of the target location are fed to the recurrent unit to model the local dynamics, the hidden representation (state) at time t is calculated as follows:

$$s_t = g(s_{t-1}, X_{t,d}^{dyn}), \tag{4.13}$$

where $g(.)$ is also LSTM, as in the global component, and s_t is the t-th hidden state of LSTM. Then, a temporal attention on top of the LSTM is also employed in this component, which can select the most informative hidden states with respect to the dynamic features of the target location d. The output of the local component ξ_d for the target location d is calculated as follows:

$$z^{(i)} = (v^{loc})^{\intercal} tanh(W^{loc} s_i + b^{loc}), \tag{4.14}$$

$$\delta^{(i)} = \frac{exp(z^{(i)})}{\sum_{t=1}^{w} exp(z^{(t)})}, \tag{4.15}$$

$$\xi_d = \sum_{t=1}^{w} \delta^{(t)} s_t, \tag{4.16}$$

where $W^{loc} \in \mathbb{R}^{m \times m}$, $b^{loc} \in \mathbb{R}^m$ and $v^{loc} \in \mathbb{R}^m$ are the parameters to be learned, and m is the number of the hidden units in LSTM in the local component. $z^{(i)}$ is the score for the contribution of the hidden state s_i. Accordingly, $\delta^{(i)}$ denotes the attention weight over the hidden state s_i.

4.2.3 Static Component

This component models the static information specific to the target location (Figure 4.1-b). The input incorporates the static features, X_d^{stat}, and a one-hot encoding vector $x_d^{id} \in \mathbb{R}^L$ that represents the target location. A fully connected layer (FC) is applied to separately learn a latent representation for each of the two types of information. In particular, the one-hot location vector will be converted into a location embedding emb_d and will be utilized in the aforementioned query component (see Eq. (4.10)). Ψ_d is the output of this component, which is concatenation of the learned location embeddings and the latent representation of the static features as follows:

$$\Psi_d = [emb_d; \sigma(W^{stat} X_d^{stat} + b^{stat})], \tag{4.17}$$

where $W^{stat} \in \mathbb{R}^{(o+p)\times(o+p)}$ and $b^{stat} \in \mathbb{R}^{o+p}$ are the network parameters to be learned in the static component. o is the number of hidden units in the FC layer for encoding the static features whereas p is the embedding dimension of the location identifier. $[;]$ is the concatenation operation. σ is the sigmoid function for the non-linear activation.

4.2.4 Objective Function

Before describing the objective function, the prediction of the number of events at the target location d at a future time t^* is computed using a linear combination of the outputs of all components as follows:

$$\hat{y}_{t^*,d} = (W_f[\Psi_d; \xi_d; \nu] + b_f), \tag{4.18}$$

where $W_f \in \mathbb{R}^{(2m+o+p)\times(2m+o+p)}$ and $b_f \in \mathbb{R}^{2m+o+p}$ are the parameters to be learned at the final fully connected layer. m is the number of hidden units in an LSTM in the local and the global components, $o + p$ is the dimension of the output of the static component. $[;]$ is the concatenation operation.

The objective function consists of three terms: the prediction loss, the orthogonality loss (constraint) and the Group Lasso regularization as follows:

$$\mathcal{L}_{overall} = \mathcal{L}_{predict} + \lambda\mathcal{L}_{ortho} + \eta\mathcal{L}_{GL}, \tag{4.19}$$

where λ and η are the hyper-parameters to be tuned for the orthogonality loss ($\mathcal{L}_{ortho}$) and the Group Lasso regularization ($\mathcal{L}_{GL}$), respectively. Moreover, $\mathcal{L}_{predict}$, which is the mean squared error (MSE), is defined as follows:

$$\mathcal{L}_{predict} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2, \tag{4.20}$$

where $\hat{y}_i$ and y_i are the predicted and the actual number of the incidents (events) for the sample i, respectively. A penalty term, $\mathcal{L}_{ortho}$ is added to avoid learning redundant memberships across communities, i.e., multiple communities may consist of a similar group of locations. To encourage community memberships to be distinguishable as much as possible, this orthogonality loss term is incorporated into the objective function. Let $\bar{\alpha}_k$ be the community membership vector denoting how each location contributes to the community k, averaging over time, and $\Delta = \left[\bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\alpha}_K\right] \in \mathbb{R}^{K \times L}$ is a matrix consisting of such membership vectors for all communities, the orthogonality loss is given by:

$$\mathcal{L}_{ortho} = \|\Delta \cdot \Delta^{\intercal} - I\|_F^2 , \tag{4.21}$$

where $I \in \mathbb{R}^{K \times K}$ is the identity matrix and $\|.\|_F^2$ denotes the squared Frobenius norm. This loss term encourages different communities to have non-identical locations as members as much as possible, which helps reduce the redundancy across communities. Lastly, the Group Lasso regularization is incorporated into the objective function, which imposes sparsity on a group level [124], and which has been found effective in several domains ([126, 138]) to select informative features. The main motivation to employ Group Lasso in the proposed method is to select community-level and local-level informative features. It enables us to interpret and differentiate which features are important for future incidents (events). It is defined as:

$$\mathcal{L}_{GL} = \sum_{k=1}^{K} \left\|Z_k^{glob}\right\|_{2,1} + \left\|Z^{local}\right\|_{2,1} + \|Z^{stat}\|_{2,1} , \tag{4.22}$$

$$\|Z\|_{2,1} = \sum_{g \in G} \sqrt{|g|}\, \|g\|_2 , \tag{4.23}$$

where Z_k^{glob} denotes the input weight matrix in the k^{th} community block in the global component. Z^{local} and Z^{stat} represent the input weight matrices in the local and the static components, respectively. g is vector of outgoing connections (weights) from an input neuron, G denotes a set of input neurons, and $|g|$ indicates the dimension of g.

4.3 Experiments

This section provides information about the datasets (including data collection process) that are used in the experiments to evaluate the proposed method, the feature sets (including feature extraction) that are employed in the model, the comparison methods for the evaluation, and the experimental settings.

4.3.1 Datasets

The proposed method CASTNet was applied to forecast opioid overdoses on two cities, namely City of Chicago and City of Cincinnati. The neighborhood boundaries officially recognized by the City of Cincinnati and the City of Chicago are called “Statistical Neighborhood Approximations (SNAs)” and “community areas”, respectively. Hereafter, the term “neighborhoods" is used to refer to both. There are 77 neighborhoods in Chicago whereas Cincinnati consists of 50 neighborhoods. While 47 neighborhoods from Chicago were selected (where $\sim$ 80% of opioid overdose deaths occurred), all neighborhoods of Cincinnati were used in the experiments. Table 4.1 shows descriptive information about both datasets. For each city, three types of data were collected related to crime, opioid overdose and the census as follows:

Table 4.1: Descriptive information about crime and opioid overdose datasets.

	#NBHDs	#Dynamic Features	#Static Features	#Crimes	#Opioid ODs	Incident Type	Time Interval	Time Span
Chicago	47	15	9	573207	1468	deaths	1 week	08/03/15 08/26/18
Cincinnati	50	10	9	75779	5401	EMS calls	1 week	08/01/15 06/01/18

4.3.1.1 Crime Data

Crime incident information including geo-location, time and primary type of the crimes was collected from the open data portals of the cities. Public Safety Crime dataset[1] and Police Data Initiative (PDI) Crime Incidents dataset[2] were used to extract such information for City of Chicago and City of Cincinnati, respectively. The crime data was collected in September 2018 and July 2018 for City of Chicago and City of Cincinnati, respectively. For the crime type, each crime incident has a hierarchical structure. Note that this hierarchical structure may exhibit different structures for the police departments of the different cities. The highest-level crime type for a crime incident was employed in the experiments. Although there exist common crime types for both cities in our datasets, there also exist different crime types. Also, note that the crime incidents whose highest-level crime type occurrences are rare during the given time period were eliminated. Besides, the dataset was cleaned in a way that the crime incidents were removed from the datasets, which are duplicates or do not have any geo-location information or time information. Finally, all the crime incidents were mapped to the corresponding neighborhoods using their geo-location information. The distributions of the total crime incidents across the neighborhoods for City of Chicago and City of Cincinnati are given in Figure A.1b and Figure A.2b, respectively. As shown in the corresponding figures, the crime incidents are not equally distributed among the neighborhoods, instead particular neighborhoods are subject to more crime occurrences.

4.3.1.2 Opioid Overdose Data

Different types of opioid overdose data were collected for each city since there is no systematic monitoring of drug abuse at either a regional-level or state-level in the United States. For City of Chicago, opioid overdose death records were collected including the geo-location and time information from Opioid Mapping Initiative Open Datasets[3]. On the other hand, the Emergency Medical Service (EMS) responses data[4] for heroin overdoses were utilized in City of Cincinnati. The opioid overdose data was collected in September 2018 and July 2018 for City of Chicago and City of Cincinnati, respectively. The overdose occurrences, which are duplicates, or which have no geo-location or time information, were removed from the datasets. In the Cith of Chicago data, the street address for each of the incidents is also provided. For the entries that have missing geo-location information, their geo-location in-

[1] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

[2] https://data.cincinnati-oh.gov/Safer-Streets/PDI-Police-Data-Initiative-Crime-Incidents/k59e-2pvf

[3] https://opioidmappinginitiative-opioidepidemic.opendata.arcgis.com/

[4] https://insights.cincinnati-oh.gov/stories/s/Heroin/dm3s-ep3u/

formation was identified from the street addresses using Geocoding API[5] of Google Maps. The instances that are false alarms were also removed from the datasets. Finally, the opioid overdose occurrences were mapped to the corresponding neighborhoods for both cities using their geo-location information. The distributions of the opioid overdose incidents across the neighborhoods for City of Chicago and City of Cincinnati are given in Figure A.1a and Figure A.2a, respectively. As shown in the corresponding figures, the opioid overdoses are not equally distributed among the neighborhoods, instead particular neighborhoods suffer from more opioid overdose incidents.

4.3.1.3 Census Data

The 2010 United States Census data, which is provided by the U.S. Census Bureau, was employed to extract the features related to demographics (population, gender distribution, race distribution), economical status, housing status and educational status. The Census data contains varying types of information from demographics to economical indicators for different spatial resolutions.

4.3.2 Features

As mentioned earlier, two types of features are incorporated into the proposed model, which are the static features and the dynamic features.

Static features include the economical status, housing status, educational level of the neighborhoods and the demographics such as population, gender diversity index and race diversity index, which were extracted from the census data. The diversity index was calculated using the normalized entropy. Since these variables (i.e. gender and race) are the sensitive variables, such an approach was followed in order to avoid making a direct conclusions on gender or race labels. Furthermore, median household income, per capita income and percentage of the poverty in a neighborhood were employed as the economical indicators. The percentage of the vacant houses (housing occupancy) and the percentage of owner occupied houses (housing tenure) were utilized as the housing-related static features. Also, the percentage of high school graduation and below was considered as the educational attainment indicator as an another static feature. As a result, a total of nine static features was obtained. Note that z-score normalization for the median household income and per capita income, and the log-transformation for population was applied while preparing the static feature vectors. As a result, nine static features were obtained for a location.

Dynamic features are to capture the crime dynamics of the locations that may be predictive for opioid overdose. the dynamic features were extracted from the public safety data portals of the cities. Each crime incident is identified by a unique crime incident number and has a certain type which shows a hierarchical structure. The crime data gathered from different cities may have very different categories. For example, the dataset from the City of Chicago includes much more categories than that from the City of Cincinnati. Here, only the highest-level, “primary crime types” were considered and the rare categories were eliminated. The full list of the crime categories used in our experiments as follows: *Part 2 Minor*,

[5] https://developers.google.com/maps/documentation/geocoding/start

Theft, *Burglary/Breaking Entering*, *Robbery*, *Aggravated Assaults*, *Rape*, *Unauthorized Use* and *Homicide* are the primary crime types for City of Cincinnati. On the other hand, *Theft*, *Battery*, *Criminal Damaging*, *Assault*, *Deceptive Practice*, *Other Offenses*, *Narcotics*, *Burglary*, *Robbery*, *Motor Vehicle Theft*, *Criminal Trespass*, *Weapons Violation* and *Homicide* are the primary crime types for City of Chicago. In addition to these features, the number of total crimes and the number of the total opioid overdose incidents were also utilized as additional dynamic features. For each neighborhood and each time unit, the feature vector contains the total number of crimes, the total number of incidents for each primary crime type and the number of opioid overdose. Zero-mean and unit variance normalization was applied to all dynamic features while preparing the feature vectors. As a result, 15 (14 crime-related, 1 opioid-related) dynamic features for City of Chicago, and 10 (9 crime-related, 1 opioid-related) dynamic features for City of Cincinnati were extracted.

4.3.3 Comparison Methods

The proposed method CASTNet was compared with several baselines and state-of-the-art approaches. In order to evaluate the forecasting effectiveness of the proposed model, two sets were selected as the comparison methods.

The first set includes Autoregressive Integrated Moving Average (ARIMA) [139], Vector Autoregerssion (VAR) [140], Support Vector Regression (SVR) since they are widely-used methods as baselines in the forecasting literature. In addition to those methods, the historical average (HA) was also included as a very basic baseline into this set. The second set of methods are the recently proposed neural network-based spatio-temporal forecasting approaches which have shown state-of-the-art performances in forecasting domain, which are DA-RNN [64], GeoMAN [65] and ActAttn [128]. In addition to these state-of-the-art methods, a basic LSTM network was also included as another neural network-based baseline method into this set. Note that none of the existing approaches either support the hierarchical structure of features obtained from intra- and inter-regions, or differentiate the inter-region contributions with respect to the target location. The comparison methods are summarized as follows:

The first set:

- *HA* is basic historical average of opioid overdose occurrences.
- *ARIMA [139]* is a well-known method for predicting future values for time series.
- *VAR [140]* captures the linear inter-dependencies among multiple time series and forecasts future values.
- *SVR* is simple Support Vector Regression. Its two variants were used, where separate models for each location were trained in SVR$_{ind}$ On the other hand, a single model for all locations was trained in SVR$_{all}$.

The second set:

- *LSTM* is a basic LSTM network. An LSTM network was trained in which the dynamic features are fed to the LSTM, then the latent representations are concatenated with the static features for the prediction.

- *DA-RNN [64]* is a dual-staged attention-based RNN model for spatio-temporal time series prediction.

- *GeoMAN [65]* is a multi-level attention-based RNN model for spatio-temporal prediction. It has shown the state-of-the-art performance in the air quality prediction task.

- *ActAttn [128]* is a hierarchical spatio-temporal predictive framework, which is proposed in the previous chapter. The final classification layer was replaced with the regression layer to configure it to the regression task and it was used as another baseline.

Furthermore, to evaluate the effectiveness of individual components of our proposed model CASTNet, its several variants were also included for the comparison as follows:

- *CASTNet-noGL*: Group Lasso regularization is not incorporated into the loss function.

- *CASTNet-noOrtho*: The orthogonality penalty is not applied so that differentiation of the communities is not encouraged.

- *CASTNet-noSA*: The spatial attentions are removed from the community blocks. Instead, the feature vectors of all locations are concatenated.

- *CASTNet-noTA*: The temporal attentions in both local and global components are removed from the architecture.

- *CASTNet-noCA*: The community attention is removed from the architecture. Instead, the context vectors of the communities are concatenated.

- *CASTNet-noSC*: The static features are excluded from the architecture, yet the location identification is still embedded to differentiate the global contributions with respect to the target location.

4.3.4 Experimental Settings

In the experiments, 'week' was used as the time unit and 'neighborhood' was used as the location unit. These units were chosen based on domain expert knowledge and availability of the data. Datasets were divided into training, validation and test sets with ratio of 75%, 10% and 15%, respectively. τ was set to 1 to make short-term predictions. For RNN-based methods, the hidden unit size of LSTMs was selected from $\{8, 16, 32, 64\}$. Also, for RNN-based methods, the experiments were performed with different window sizes $w \in \{5, 10, 15, 20\}$. The networks were trained using Adam optimizer [135] with a learning rate of 0.001. For each LSTM layer, dropout of 0.1 was applied to prevent overfitting. In the proposed models, the regularization factors λ and η were optimized from the small sets $\{0.001, 0.005, \ldots, 0.05\}$ and $\{0.001, 0.0015, \ldots, 0.01\}$, respectively using grid search. For *ARIMA* and *VAR*, the orders of the autoregressive and the moving average components were optimized for the time lags between 1 and 11. Finally, the code for proposed method and the data are available at `https://github.com/picsolab/castnet`.

4.4 Results

In this section, a comprehensive set of results is presented. First, in Section 4.4.1, the forecasting effectiveness of the proposed model in comparison with the baseline and state-of-the-art forecasting approaches, and based on the aforementioned experiment settings is shown. In Section 4.4.2, different kinds of spatio-temporal contributions (community memberships, community contributions and temporal attentions) of the proposed model are analyzed and interpreted. In Section 4.4.3, we analyze different kinds of predictive features identified by the proposed model are analyzed and their effect on the prediction is interpreted.

4.4.1 Performance Comparison

The forecasting performance of CASTNet was compared with the comparison methods. The results are organized to answer the following three questions:

1. Overall, how well could CASTNet forecast future opioid overdoses, compared with the baseline methods? (Section 4.4.1.1)
2. How well could CASTNet forecast future opioid overdoses in individual-level (for each location), compared with the baseline methods? (Section 4.4.1.2)
3. How forecasting performance of CASTNet does change with respect to the length of the window size, compared with the baseline methods? (Section 4.4.1.3)

4.4.1.1 Overall Performance

Table 4.2 shows that CASTNet achieves the best performance in terms of both mean absolute error (MAE) and root mean squared error (RMSE) on both datasets. The model shows 17.2% and 5.3% improvement in terms of MAE and RMSE, respectively, on Chicago dataset compared to state-of-the-art approach GeoMAN. Similarly, CASTNet enhances the performance 6.3% and 2.4% on Cincinnati dataset in terms of MAE and RMSE, respectively, compared to DA-RNN which shows best performance among the other baselines. Furthermore, it is observed that mostly spatio-temporal RNN-based models outperform other baselines, which indicates they better learn the complex spatio-temporal relationships between crime and opioid overdose dynamics.

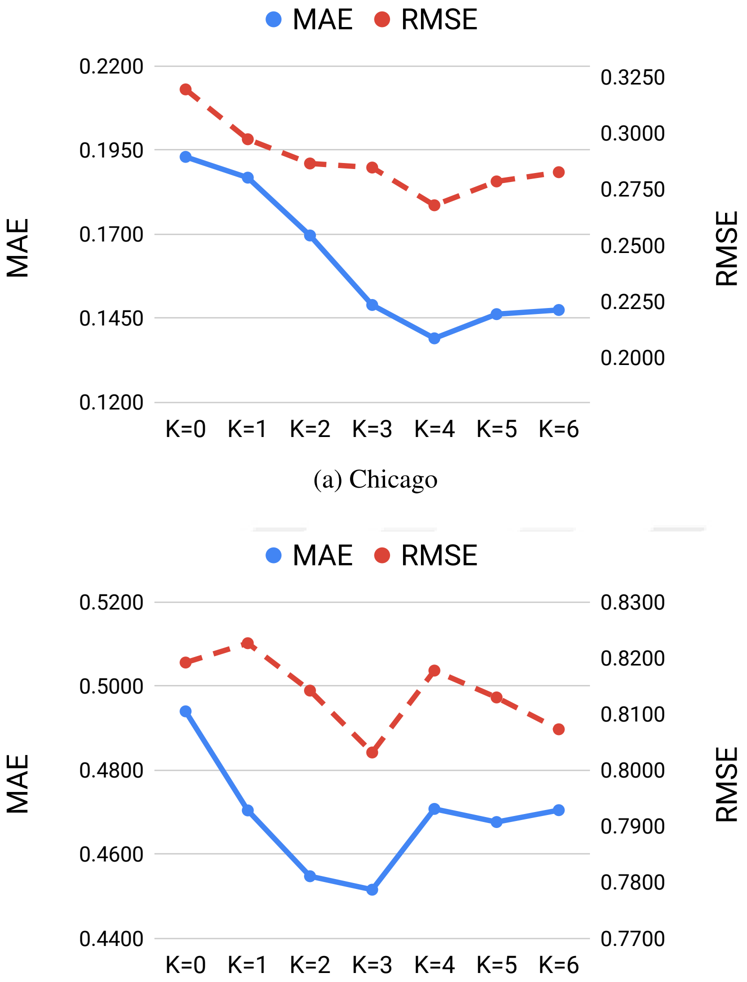
The effectiveness of each individual component of CASTNet is further evaluated with an ablation study. As described in Section 4.3.3, each variant is different from the proposed CASTNet by removing one tested component (with others kept identical as much as possible). Table 4.2 shows that the removal of Group Lasso regularization from the loss function results in a significantly lower performance compared to the others. In addition, CASTNet-noGL can no longer be able to select informative features loses its capability of interpretability for important features. Similarly, excluding the orthogonality loss term (CASTNet-noOrtho) results in losing the ability to learn distinguishable communities or representation subspaces, and reduces the forecasting performances as well. Moreover, comparing CASTNet with CASTNet-noCA indicates that employing community attention has a great impact on the forecasting

Table 4.2: Forecasting performance results.

	Chicago		Cincinnati	
	MAE	RMSE	MAE	RMSE
HA	0.2329	0.3385	0.5728	0.8727
ARIMA [139]	0.2272	0.3396	0.5717	0.8952
VAR [140]	0.2242	0.3386	0.5606	0.8712
SVR$_{ind}$	0.2112	0.3321	0.5153	0.8609
SVR$_{all}$	0.1984	0.3063	0.4886	0.8602
LSTM	0.2024	0.3134	0.5235	0.8267
DA-RNN [64]	0.1726	0.3051	0.4817	0.8225
GeoMAN [65]	0.1679	0.2829	0.5034	0.8453
ActAttn [128]	0.1693	0.2937	0.4827	0.8326
CASTNet-noGL	0.1662	0.3129	0.4703	0.8311
CASTNet-noOrtho	0.1649	0.2948	0.4716	0.8109
CASTNet-noSA	0.1608	0.2893	0.4579	0.8152
CASTNet-noTA	0.1641	0.2876	0.4700	0.8141
CASTNet-noCA	0.1631	0.3069	0.4730	0.8225
CASTNet-noSC	0.1693	0.2980	0.4692	0.8291
CASTNet	**0.1391**	**0.2679**	**0.4516**	**0.8032**

performance, which indicates that learning pairwise activity relationships between a particular event location and the communities is crucial. Location-specific static features are also informative since their exclusion (CASTNet-noSC) degrades the performance in both cases. Furthermore, removal of temporal attentions in both local and global components (CASTNet-noTA) results in a decrease in the forecasting performance as well as loss of interpretibility in important previous time steps for forecasting future opioid overdoses. Last but not the least, the individual component that provides the least performance gain is spatial attention for both cases. However, its removal (CASTNet-noSA) results in loss of interpretability capability of community memberships. These results reflect that each individual component has important contribution to forecasting performance.

Moreover, the forecasting performance of the CASTNet is evaluated with respect to the change in the number of communities K. Experiments with different values of K selected from $\{0, 1, \ldots, 6\}$ are conducted, and the results are given in Figure 4.2. Note that the model does not consider global contribution when $K = 0$. Also, when $K = 1$, the model yields a single universal representation of global activities which is irrespective of the event locations. The best performances are obtained when $K = 4$ for Chicago and $K = 3$ for Cincinnati datasets. It is observed that while K increases until the optimum value, the performance increases, and some communities are decomposed to form new communities. However, as long as K continues to increase after its optimum value, the performance starts to decrease slightly or remains stable, and the semantic subspaces of some communities become similar. With this experiment, it is indicated that instead of learning a fixed and single universal representation of global activities, learning different representations subspaces for the global activities significantly improves the forecasting performance.



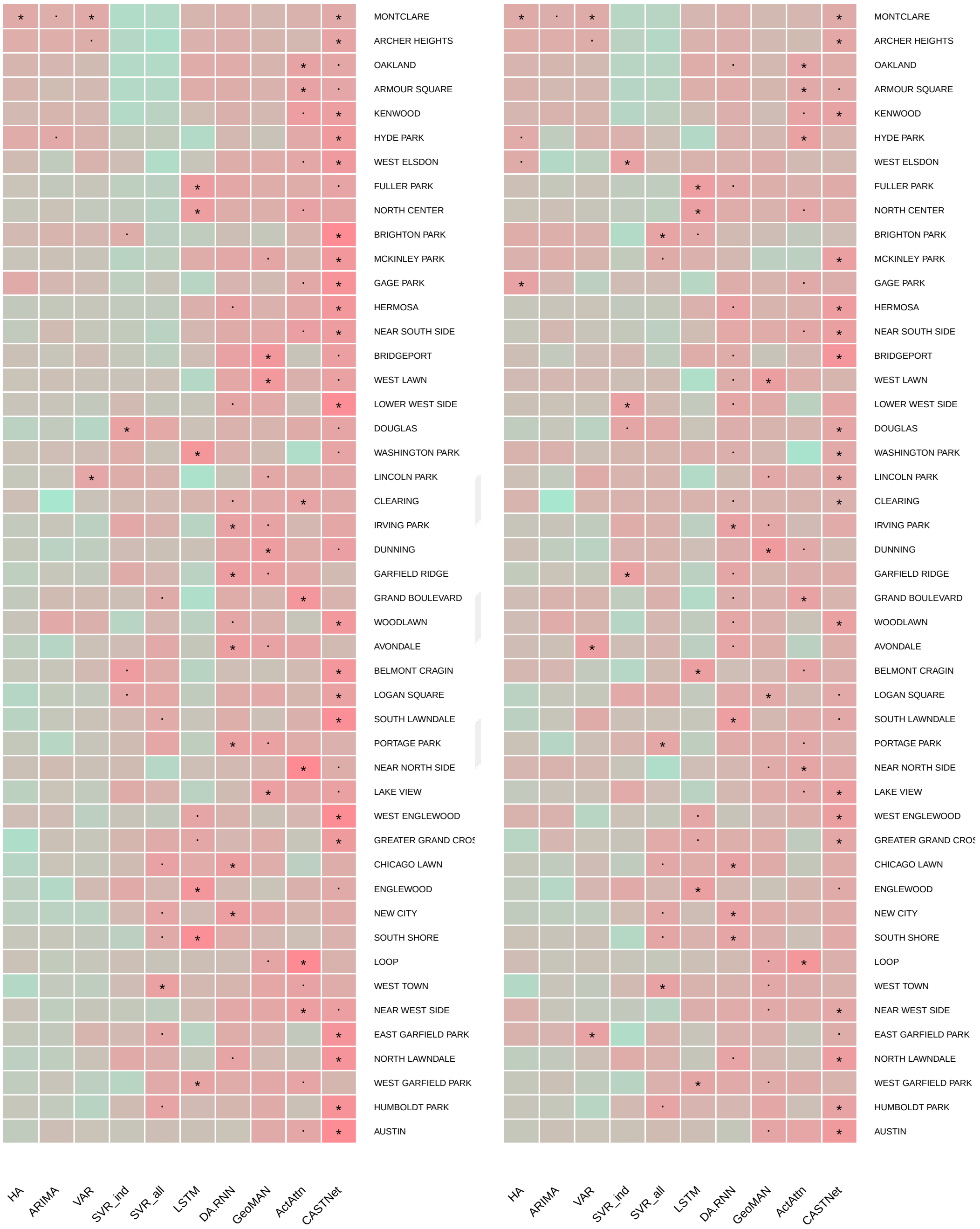
(a) Chicago

(b) Cincinnati

Figure 4.2: MAE and RMSE results w.r.t change in the number of communities.

4.4.1.2 Individual-level Performance Analysis

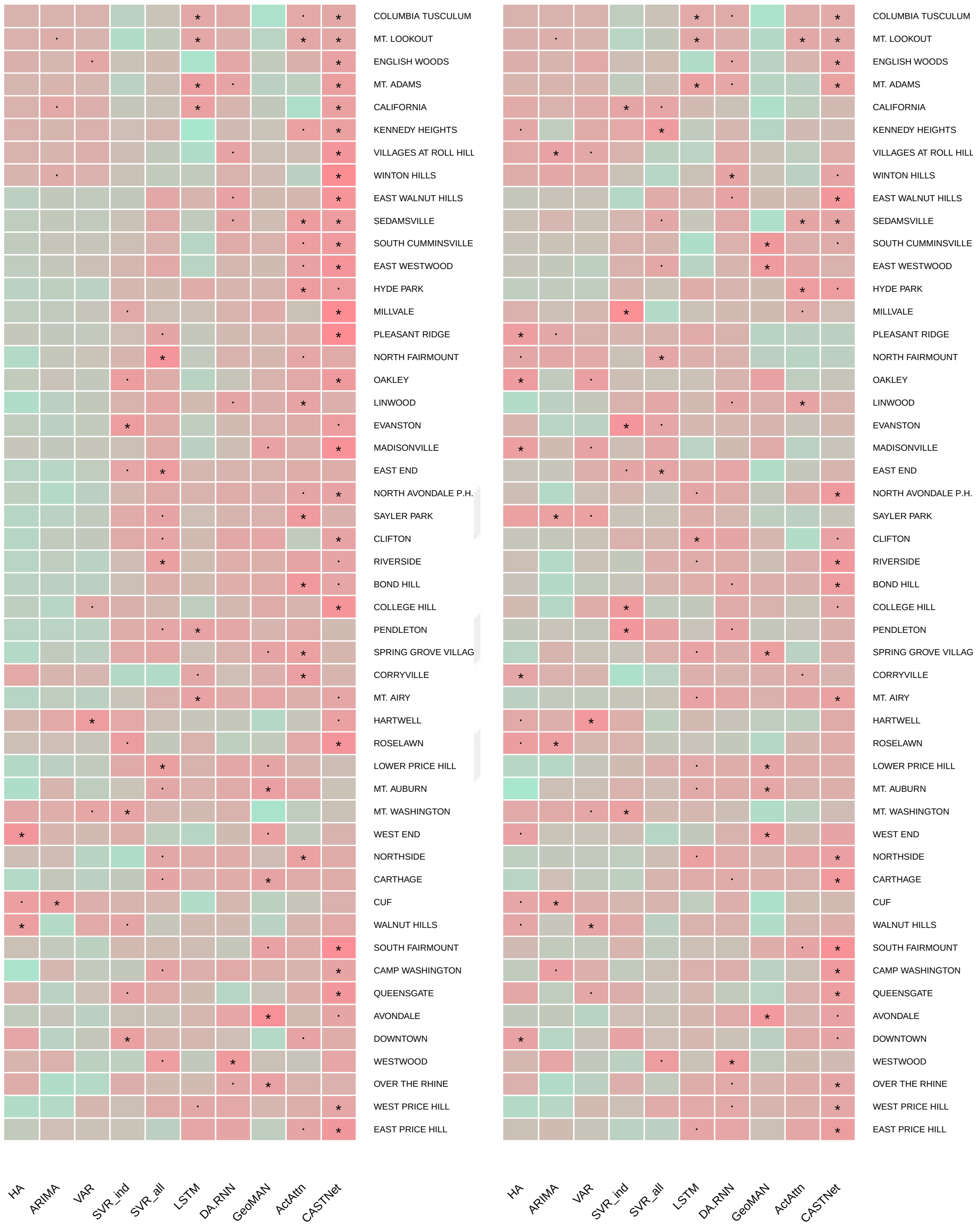
In addition to evaluating the overall forecasting performance of CASTNet, in-depth performance analysis is also performed by investigating individual-level performances. In other words, the performance of CASTNet is compared with the baselines for each of the neighborhoods. Figure 4.3 and Figure 4.4 indicate the individual-level forecasting performances of all methods via heatmaps for City of Chicago and City of Cincinnati, respectively. Each row in the figures reveals the errors of the methods for the associated neighborhood through a color scale from red to green. The closer the color to red, the less the error. On the other hand, the closer the color to green, the more the error. While "$*$" in a cell indicates the method with the lowest error (the best performing method) for the given neighborhood, "·" in the cell denotes the runner up for the same neighborhood. Note that the neighborhoods in both figures are ordered by the number of total opioid overdoses occurred in ascending order from top to down.



(a) MAE in individual-level.

(b) RMSE in individual-level.

Figure 4.3: Error heatmaps for (a) MAE and (b) RMSE in individual-level (neighborhood-level) for City of Chicago. While y axis represents the neighborhoods, x axis indicates the methods. The closer the color to the red, the less the corresponding error. On the other hand, the closer the color to the green, the more the corresponding error. While "$*$" in a cell indicates the method with the lowest error (the best performing method) for the given neighborhood, "$\cdot$" in the cell denotes the runner up for the same neighborhood.



(a) MAE in individual-level.

(b) RMSE in individual-level.

Figure 4.4: Error heatmaps for (a) MAE and (b) RMSE in individual-level (neighborhood-level) for City of Cincinnati. While y axis represents the neighborhoods, x axis indicates the methods. The closer the color to the red, the less the corresponding error. On the other hand, the closer the color to the green, the more the corresponding error. While “$*$” in a cell indicates the method with the lowest error (the best performing method) for the given neighborhood, “$\cdot$” in the cell denotes the runner up for the same neighborhood.

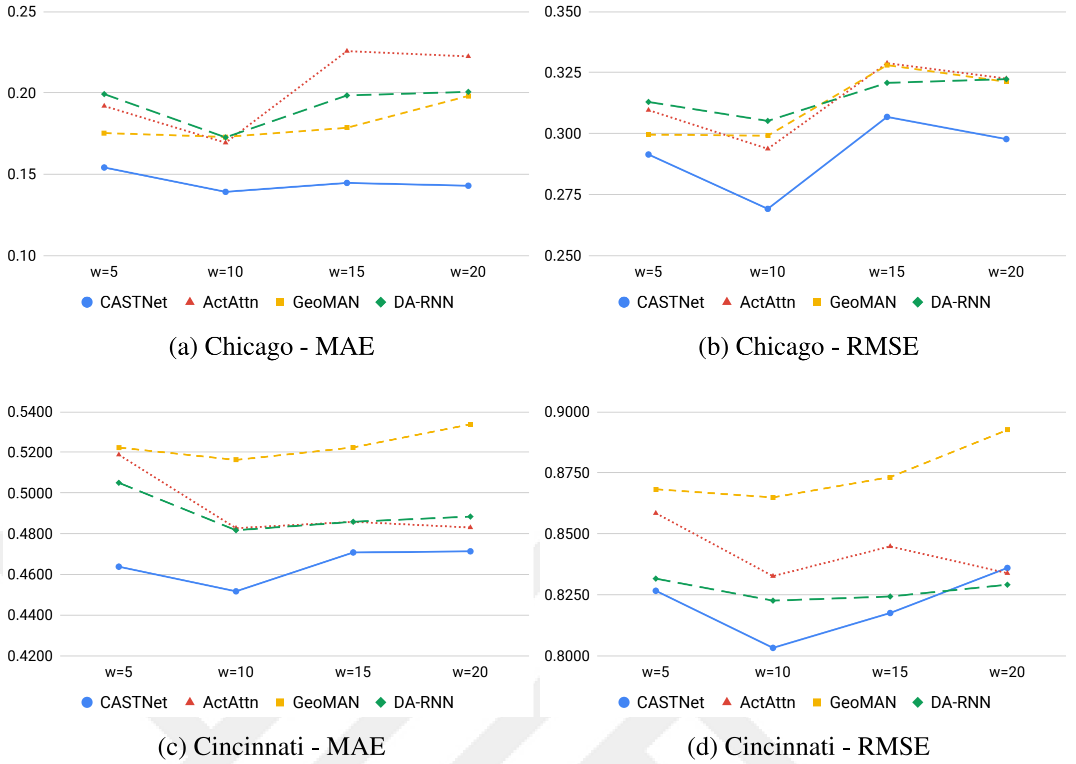
Table 4.3: Significance test for the evaluation by Wilcoxon test. $*$ is p < 0.05, and $**$ is p < 0.0056 (which is p after Bonferroni correction due to multiple comparisons.).

		HA	ARIMA	VAR	SVR$_{ind}$	SVR$_{all}$	LSTM	DA-RNN	GeoMAN	ActAttn
Chicago	MAE	3.6e-09**	1.4e-14**	3.9e-09**	1.6e-12**	5.4e-08**	4.8e-07**	0.0003**	0.0004**	0.0015**
	RMSE	1.3e-08**	3.6e-13**	2.7e-08**	1.2e-10**	4.4e-06**	2.8e-06**	0.0117*	0.0010**	0.0008**
Cincinnati	MAE	3.2e-08**	8.7e-09**	3.7e-09**	6.8e-06**	0.0005**	8.8e-07**	0.0004**	8.9e-05**	0.0034**
	RMSE	0.0011**	3.2e-05**	0.0001**	0.0002**	8.9e-05**	3.5e-05**	0.0111*	2.8e-06**	9.4e-07**

For City of Chicago (Figure 4.3a), the proposed method CASTNet performs the best among the others in 21 neighborhoods, and the second among the others in 12 neighborhoods out of 47 neighborhoods with respect to MAE. The runner-up is GeoMAN (according to the overall performance), and it performs the best in 4 neighborhoods and the second in 7 neighborhoods. It can also be considered that ActAttn is the runner-up based on the performance in individual-level, which achieves the best performance in 7 neighborhoods and the second performance in 8 neighborhoods. Also, as given in Figure 4.3b CASTNet performs the best among the others in 19 neighborhoods, and the second among the others in 5 neighborhoods out of 47 neighborhoods with respect to RMSE. GeoMAN performs the best in 3 neighborhoods and the second in 8 neighborhoods with respect to RMSE. ActAttn achieves the best performance in 6 neighborhoods and the second performance in 8 neighborhoods with respect to RMSE. It is also observed that CASTNet performs well in the most of the neighborhoods of City of Chicago regardless of the number of opioid overdoses occurred in these neighborhoods.

Furthermore, for City of Cincinnati (Figure 4.4a), the proposed method CASTNet performs the best among the others in 24 neighborhoods, and the second among the others in 7 neighborhoods out of 50 neighborhoods with respect to MAE. The runner-up method DA-RNN (according to the overall performance) yields the best performance among the other methods only in 1 neighborhood and performed the second in 6 neighborhoods. ActAttn can also be considered as the runner-up based on the performance in individual-level, which achieves the best performance in 6 neighborhoods and the second performance in 7 neighborhoods. Also, as given in Figure 4.4b CASTNet performs the best among the others in 18 neighborhoods, and the second among the others in 7 neighborhoods out of 50 neighborhoods with respect to RMSE. Da-RNN performs the best in 2 neighborhoods and the second in 10 neighborhoods with respect to RMSE. ActAttn achieves the best performance in 4 neighborhoods and the second performance in 3 neighborhoods with respect to RMSE. It is also observed that CASTNet performs well in the most of the neighborhoods of City of Cincinnati regardless of the number of opioid overdoses occurred in these neighborhoods.

A significance test is further performed to evaluate the success of the proposed method. Since there are nine baselines, nine hypotheses should be tested (CASTNet vs. nine other baselines). Table 4.3 presents the corresponding significance test results performed by Wilcoxon test with respect to the significance levels (p-value < 0.05) and (p-value < 0.0056) after Bonferroni correction due to multiple comparisons. For each error type (MAE and RMSE) and for each dataset, two distributions of individual-level errors (CASTNet vs. the baseline) are compared. As given in Table 4.3, the proposed method significantly outperforms almost all other base-



(a) Chicago - MAE

(b) Chicago - RMSE

(c) Cincinnati - MAE

(d) Cincinnati - RMSE

Figure 4.5: MAE and RMSE results with respect to different length of window size w. While x axis represents the length of the window size w, y axis indicates the corresponding error.

line methods in terms of both MAE and RMSE even after Bonferroni correction. The only comparisons (CASTNet vs. DA-RNN) for RMSE and for both datasets are not significant after Bonferroni correction, yet they are still significant before Bonferroni correction. The reason behind this situation may be the fact that both models are affected by the outliers in the datasets similarly at some degree since RMSE is more sensitive metric to outliers compared to MAE.

4.4.1.3 Performance Analysis with Varying Length of Window Size

The results shared for the forecasting performance in the previous sections are based on the best window size setting for all methods, which is $w = 10$. Additional experiments are also conducted to analyze the forecasting performance of the RNN-based state-of-the-art methods with respect to the different length of the window size where $w \in \{5, 10, 15, 20\}$. The performance results of those methods are provided for ActAttn [128], GeoMAN [65] and DA-RNN [64]) as well as the proposed method CASTNet, with different length of window size w settings in Figure 4.5. Based on these results, the proposed method CASTNet achieves superior forecasting performance in terms of MAE for all window size settings on both datasets. Also, it significantly performs better than the other baseline methods in terms of RMSE at almost all cases for both datasets. It is also observed that the performances of the methods increase until the optimum length of window size ($w = 10$), then they start decreasing as the length of the window size increases.

4.4.2 Analysis of Community Memberships and Community Contributions

In this section, the learned communities, community memberships of the neighborhoods and community contributions on forecasting future opioid overdoses are analyzed. In other words, interpretations about future opioid overdoses through the hierarchical attention mechanism of the proposed model (see Figure 4.1-c and Figure 4.1-d) are provided by answering the following questions:

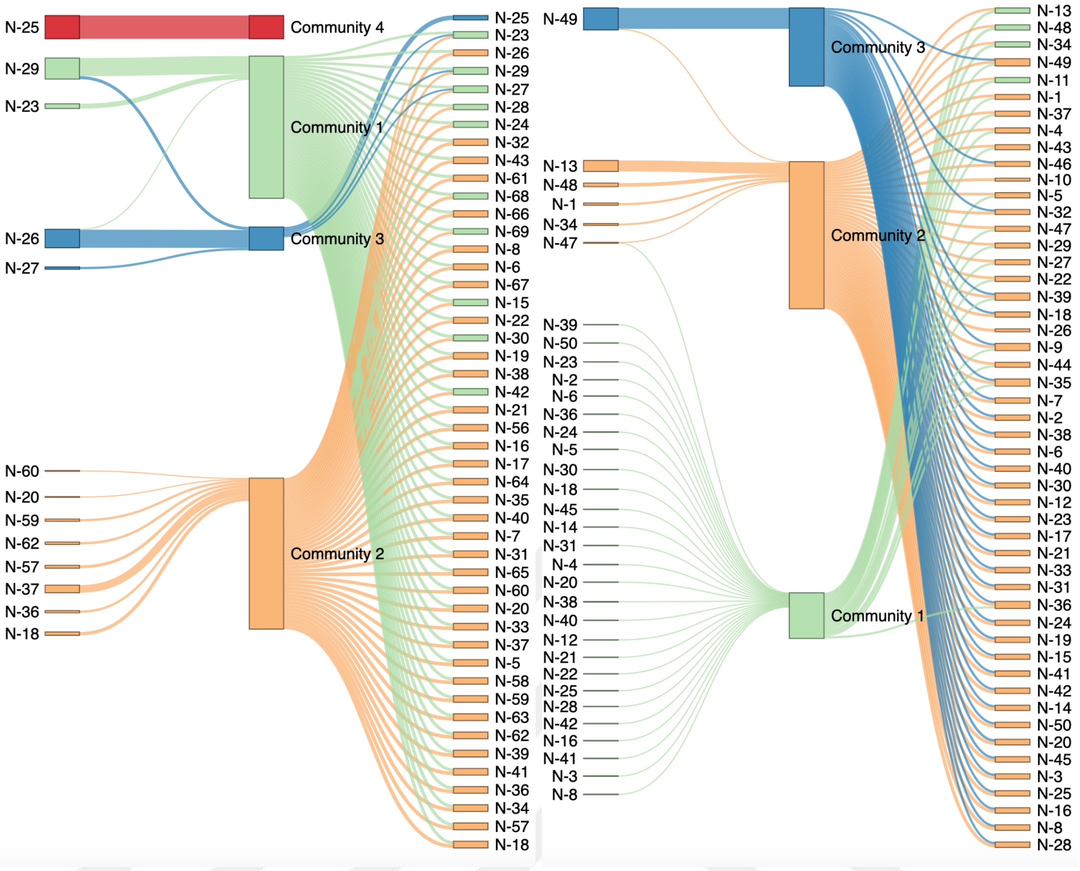
4.4.2.1 How do locations contribute to communities?

CASTNet learns different representation subspaces (communities) of global dynamic features unlike the previous work [65, 128], and each community is encouraged to consist of a group of different members due to the orthogonality penalty. In the following analyses, the results are presented in two ways; (i) by aggregating from all test samples for ease of interpretation (ii) by selecting the representative test samples.

First, the learned communities and their memberships (i.e. the spatial attention weights α in Eq. (4.2)), averaged over time through the test samples, are depicted on the left sides of Figure 4.6a and Figure 4.6b for City of Chicago and City of Cincinnati, respectively. The purpose is to observe the general behaviors of the neighborhoods as the members. The line thickness at the left sides of the figures represents the degree at which a location contributes to the corresponding communities. Note that the neighborhoods at the left sides of Figure 4.6a and Figure 4.6b are ordered by the number of crimes incidents committed in those neighborhoods. As shown in Figure 4.6, most of the neighborhoods have dedicated to only a single community, which indicates the effect of orthogonality penalty in the loss function.

For Chicago model (Figure 4.6a), *Austin (N-25)*, which has the highest number of crime incidents and opioid overdose deaths, formed a separate community C_4 by itself. While *North Lawndale (N-29) and Humboldt Park (N-23)* together formed the community C_1, *West Garfield Park (N-26)*, *East Garfield Park (N-27)* and *North Lawndale (N-29)* formed another community C_3. Note that the neighborhoods of C_1 and C_3 have the highest opioid overdose death rate after *Austin (N-25)*. On the other hand, the community C_2 was formed by the neighborhoods having low crime incident and overdose death rates including *Fuller Park (N-37), McKinley Park (N-59)* and *West Elsdon (N-62)*. Furthermore, for Cincinnati model (Figure 4.6b), *Westwood (N-49)*, where the highest number of crimes were committed, formed a separate community C_3 by itself. It shows a similar behavior to the Chicago case. *East Price Hill (N-13), West Price Hill (N-48), Avondale (N-1)* and *Over-The-Rhine (N-34)* formed the community C_2 where these neighborhoods have the highest crime rate after *Westwood (N-49)* and the highest opioid overdose rate. On the other hand, the community C_1 was formed by rest of the neighborhoods (with low and moderate crime rates) and their memberships to that community are almost equal.

Furthermore, the community memberships are investigated while forecasting the number of opioid overdoses for a specific time step. More specifically, the spatial attention weights (i.e. α in Eq. (4.2)) are demonstrated in Figure 4.7 and Figure 4.8 for City of Chicago and City of Cincinnati, respectively, while forecasting the number of opioid overdoses at the 6^{th} test week as a representative time step. Note that the spatial attention weights do not change



(a) Chicago

(b) Cincinnati

Figure 4.6: Community memberships and community contributions on forecasting future opioid overdoses. For each community, the left side represents community memberships (how each location contributes to the community), and the right side represents the average community contribution (how the community contribute to predicting a target location). The thickness of the edges at indicates the weight of community membership at the left sides of the figures while it corresponds to community contribution at the right sides of the figures. Node size denotes the overall community membership of a location (left side) and overall community contribution to forecasting overdose (right side) in the target neighborhood. Edge color shows the input and output of a specific community. Node color of a neighborhood indicates the community for which the corresponding neighborhood has the highest membership (left side). Node color of a neighborhood denotes the community from which the neighborhood takes the largest contribution (right side). Edges whose weights are above a certain threshold are shown for the sake of clarity.

from neighborhood to neighborhood since the proposed model starts differentiating target location specific global contribution at the temporal attention in the global component. In the figures, while the x axis shows the neighborhoods, the y axis indicates the time units. A green cell represents the contribution of dynamic features of the corresponding neighborhood at a specific time in the corresponding community. The darker green color denotes more contribution by the neighborhood to the corresponding community. For Chicago case (see Figure 4.7), it is observed that the most contributing neighborhoods in C_1, C_2 and C_3 change from time to time. For instance, *North Lawndale (N-29)* is the most contributing neighborhood at time $t \in \{1, 2, 3, 6, 8, 9\}$ whereas *Humboldt Park (N-23)* is the most dominant one at time $t \in \{4, 5, 7, 10\}$. *Austin (N-25)* is the only neighborhood contributing to C_4. Similarly, for Cincinnati case, the most contributing neighborhoods to forecasting in C_1 and C_2 change from time to time. For instance, *East Price Hill (N-13)* is the most contributing neighborhood at time $t \in \{1, 3, 5, 8, 9, 10\}$ for the C_2. *West Price Hill (N-48)* is the most dominant one at time $t = 7$, *Avondale (N-1)* is the most contributing neighborhood at time $t \in \{2, 6\}$. *Westwood (N-49)* is almost the only neighborhood contributing to C_3 and it always contributes the most to C_3 than any other neighborhoods. Finally, the geomaps showing the community memberships of the neighborhoods are also shown in Figure B.1a and Figure B.1b for Chicago model and Cincinnati model, respectively, in Appendix B.

4.4.2.2 How do the communities contribute to forecasting?

CASTNet is capable of modeling the pairwise activity relationships between a particular event location and the communities. It allows the target location to attend the communities to select location-specific global contributions to forecast local incidents. How these communities contribute to forecasting is analyzed (i) by visualizing the community attention weights (i.e. γ in Eq. (4.11)) averaged over test samples for each neighborhood for ease of interpretation (to summarize the community contributions for each neighborhood) in Figure 4.6a and Figure 4.6b for Chicago and Cincinnati, respectively, (ii) by inspecting representative test samples from each model.

First, the community contributions, averaged over time through the test samples, are represented on the right sides of Figure 4.6a and Figure 4.6b for City of Chicago and City of Cincinnati, respectively. It provides information about the summarized community contributions to each neighborhood. Note that the neighborhoods on the right side of Figure 4.6a and Figure 4.6b are ordered by the number of opioid overdose incidents. For Chicago case, C_1 and C_2 have more contributions than the other communities on forecasting overdose. While C_2 contributes more to neighborhoods with low or moderate opioid overdose death rate, C_1 contributes more to the neighborhoods where the death rate is higher. C_3 also contributes more to the neighborhoods with the highest death rate (e.g. *Austin (N-25), Humboldt Park (N-23)*). This means that any particular neighborhood attends more to the community, which is formed by the similar neighborhoods. On the other hand, C_4 does not significantly contribute to any neighborhood although it is formed by a crime hot-spot (*Austin (N-25)*). Moreover, for Cincinnati case, C_2 is a very dominant community, which makes the largest global contribution to most of the neighborhoods. The neighborhoods that formed C_2 and C_3 (e.g. *East Price Hill (N-13), West Price Hill (N-48), Westwood (N-49)*) are very predictive, and the change in their dynamics have greater impact on forecasting future overdoses in the target neighborhoods. On the other hand, C_1 has larger contribution to neighborhoods where the overdose rate is the highest. This indicates that the crimes committed in the members of C_1 are also informative for forecasting future overdoses in opioid hot-spots.

Second, the community contributions are investigated while forecasting the number of overdoses for a specific neighborhood at a specific time. The community attention weights (γ in Eq. (4.11)) are demonstrated during opioid overdose prediction for *North Lawndale (N-29)* from City of Chicago and for *Avondale (N-1)* from City of Cincinnati at the 6^{th} test week.

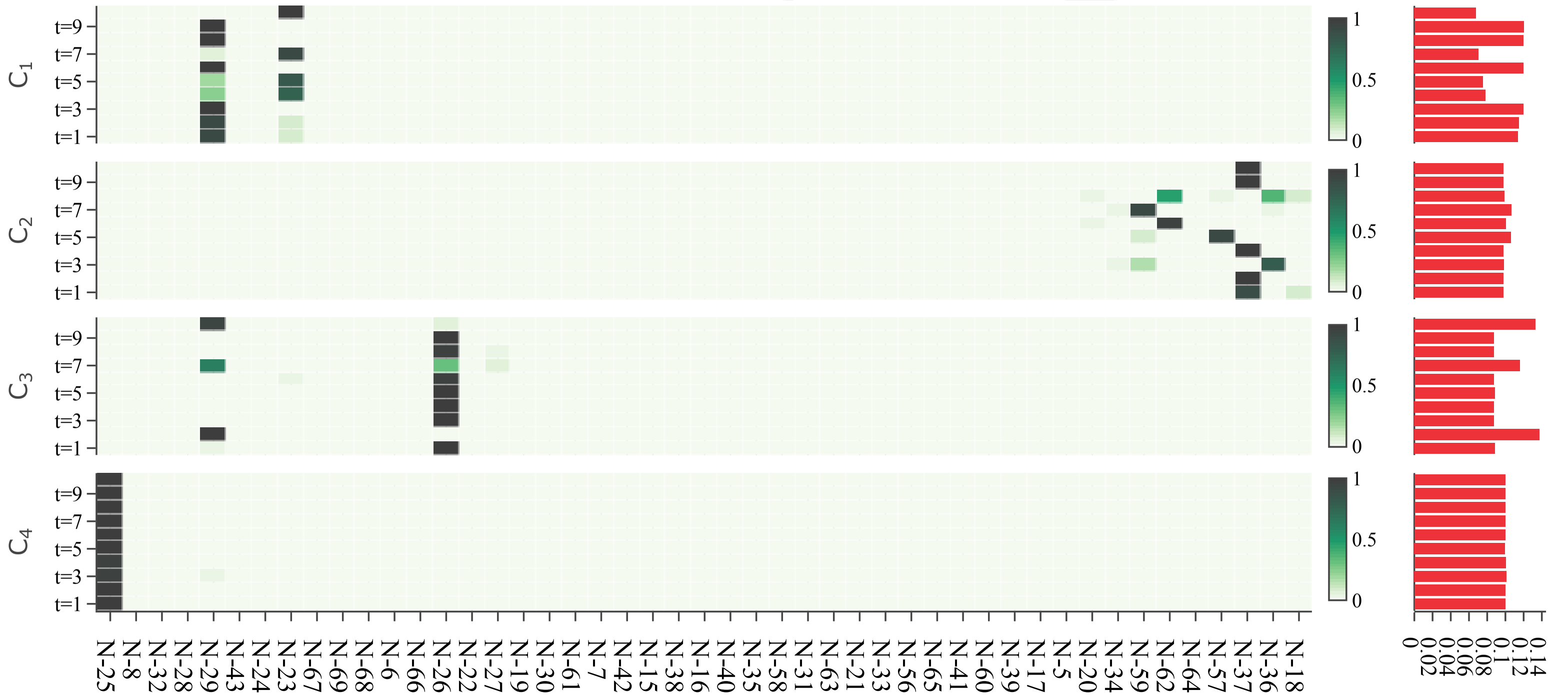


Figure 4.7: Community memberships (i.e. the spatial attention weights α in Eq. (4.2)) and temporal attention weights (β in Eq. (4.6)) of the Chicago model during the prediction for *North Lawndale (N-29)* at the 6^{th} test week. While the x axis shows the neighborhoods, the y axis indicates the time units. A green cell represents the contribution of dynamic features of the corresponding neighborhood (i.e. its spatial attention weight) at a specific time in the corresponding community. The darker green color denotes the more contribution by the corresponding neighborhood. Red bars indicate the magnitude of the temporal attention weights for each community across time.

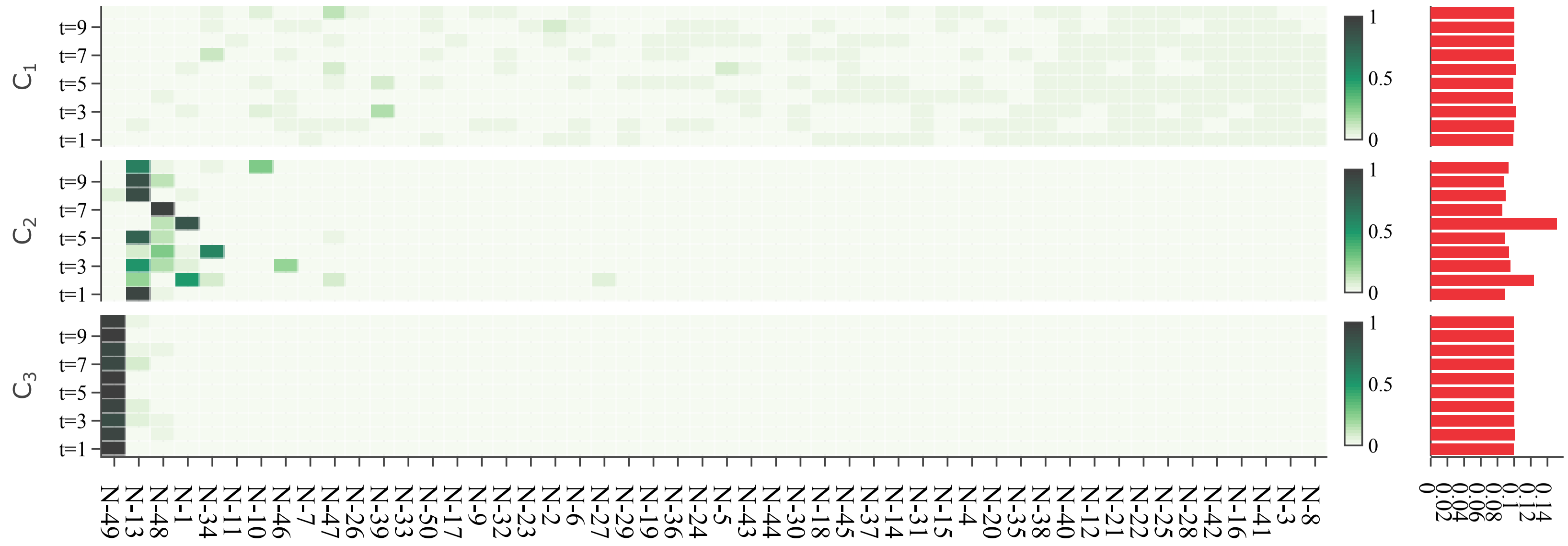


Figure 4.8: Community memberships (i.e. the spatial attention weights α in Eq. (4.2)) and temporal attention weights (β in Eq. (4.6)) of the Cincinnati model during the prediction for *Avondale (N-1)* at the 6^{th} test week. While the x axis shows the neighborhoods, the y axis indicates the time units. A green cell represents the contribution of dynamic features of the corresponding neighborhood (i.e. its spatial attention weight) at a specific time in the corresponding community. The darker green color denotes the more contribution by the corresponding neighborhood. Red bars indicate the magnitude of the temporal attention weights for each community across time.

For the sample from Chicago, the community attention weights are 0.245 for C_1, 0.299 for C_2, 0.312 for C_3, and 0.144 for C_4. On the contrary to summary for this neighborhood (i.e. C_1 contributes most to *North Lawndale (N-29)* overall, see Figure 4.6a), C_3 is the most contributing community to forecast the number of local overdose incidents in *North Lawndale (N-29)* at the 6^{th} test week. Moreover, for Cincinnati case, the community attention weights are 0.31 for C_1, 0.383 for C_2 and 0.307 for C_3. C_2 is the most contributing community during the prediction for *Avondale (N-1)* at the 6^{th} week, which is also consistent with the overall contribution to this neighborhood (see Figure 4.6b).

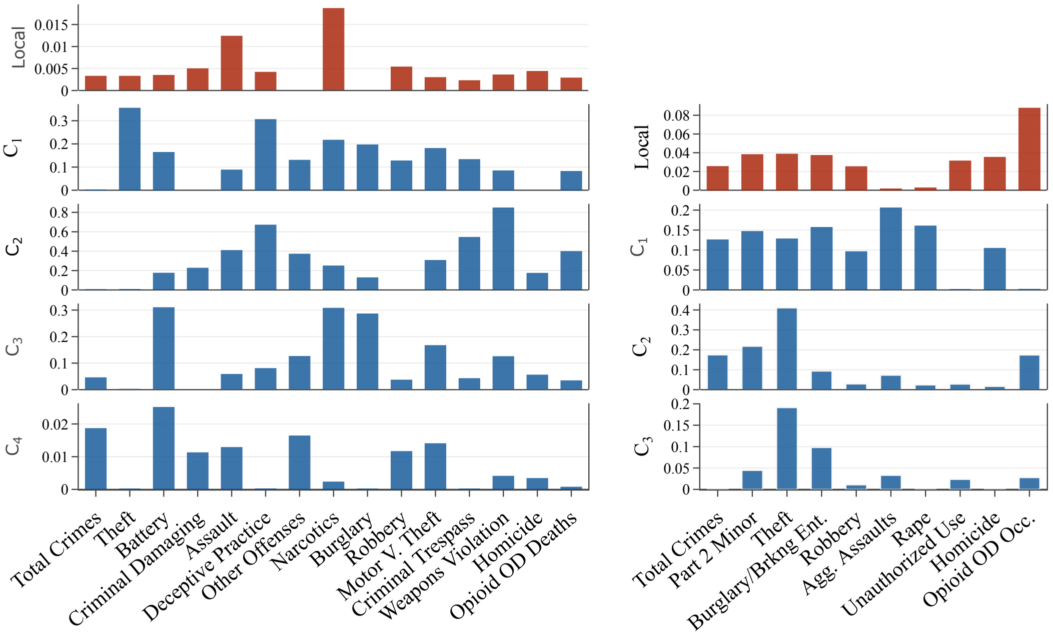
4.4.2.3 How do the community contributions change over time?

CASTNet enables us to analyze the individual community contributions across time during the prediction. Recall that the final contribution of a community for the prediction of a location at a future time is the weighted sum of its contributions over time. The time steps in which nearby locations are more active get larger weights than the others (see Eq. (4.5)). To analyze the contributions of each individual communities over time, he temporal attention weights (i.e. β in Eq. (4.6)) are analyzed for the representative test samples. More specifically, the temporal attention weights are demonstrated during the prediction of opioid overdoses for *North Lawndale (N-29)* from City of Chicago and for *Avondale (N-1)* from City of Cincinnati at the 6^{th} test week in Figure 4.7 and Figure 4.8, respectively. The right sides of the corresponding figures (red bars) indicate the temporal attention weights in each time step.

For Chicago model, it is observed that there is no a significant change in contributions over time for C_4 since this community is strongly dominated by only a single neighborhood, *Austin (N-25)*. C_2 and C_3 are the most contributing communities to forecasting local overdoses for this test sample. $t \in \{5, 7\}$ contributes slightly more than other time steps in C_2 since the most dominant membership at these time steps is by *Archer Heights (N-57)* and McKinley *Park (N-59)*, respectively, since they are spatially closer neighborhoods to the target neighborhoods than the other members of this community. Similarly, $t \in \{3, 7, 10\}$ have more contribution than other time steps for C_3 as the most contributing member at these time steps is the target location itself (*North Lawndale (N-29)*). The same behavior is observed for C_1 as well. Furthermore, for Cincinnati case, the contributions do not exhibit a different behavior over time for C_3 since it is already formed by a single neighborhood *Westwood (N-49)*. The time steps $t \in \{2, 6\}$ contribute more to the final community contribution as the most contributing member of C_3 at these time steps is the target location itself (*Avondale (N-1)*). On the other hand, any significant change is not observed in the contributions of C_1. Most of the neighborhoods are the members of C_1 with nearly equal membership at all time steps, and this may be the reason why the community contribution is almost the same for all time steps.

4.4.3 Feature Analysis

The importance of **dynamic features** is investigated by analyzing the mean absolute input weights of local and global components as shown in Fig. 4.9. For Chicago case, Group Lasso selects *Narcotics* and *Assault* as the most important features for future opioid overdose deaths in the same location. Moreover, *Theft, Deceptive Practice, Narcotics, Burglary* and *Motor V. Theft* are the predictive features from C_1 while *Weapons Violation, Deceptive Practice (e.g.*

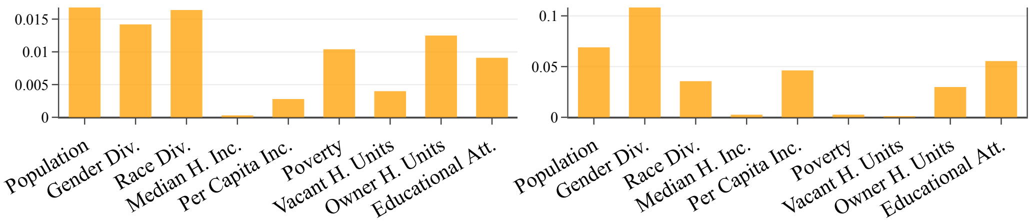


(a) Chicago

(b) Cincinnati

Figure 4.9: Importance of the dynamic features. Mean absolute values of input weights of the local and the global components.

Fraud) and *Criminal Trespass* are significant from C_2. Recall that, C_1 and C_2 are the most contributing communities to forecasting future overdoses (see Figure 4.6a). This shows that property crimes (*e.g. Theft, Burglary, Deceptive Practice*) are more significant predictors than the violent crimes for Chicago. Such crimes previously committed in the members of C_1 and C_2 (*North Lawndale (N-29), Humboldt Park (N-23), Fuller Park (N-37), etc.*) may be a significant indicator of future opioid overdose deaths in City of Chicago. On the other hand, *Battery, Narcotics, Burglary, and Motor V. Theft* are predictive features from C_3 while *Battery, Total Crimes* and *Other Offenses (e.g. offenses against family)* are significant from C_4. However, C_3 has larger contribution than other communities for only *Austin (25)*. C_4 does not provide a significant contribution to any neighborhood as much as the other communities. For Cincinnati case, *Opioid Overdose Occ.* is the most predictive feature for forecasting future opioid overdose in the same location, which means the local component behaves as an autoregressive module unlike the Chicago case. Furthermore, both violent crimes including *Agg. Assaults, Rape, Homicide, Part 2 Minor (e.g. Menacing)* and property crimes including *Burglary/Breaking Ent., Theft, Part 2 Minor (e.g. Fraud)* are significant features from C_1. On the other hand, *Theft* and *Part 2 Minor* from C_2, and *Theft* and *Burglary* from C_3 are predictive features for future opioid overdose in the target locations. Recall that C_2 and C_3 have more salient contribution on most of the neighborhoods, which implies that commitment of previous property crimes (especially *Theft*) in the members of those communities (*East Price Hill (13), West Price Hill (48), Over-The-Rhine (34), Westwood (49)* and *Avondale (1)*) may be one of the potential indicators of future opioid overdoses in the other neighborhoods. Note that these findings are also consistent with the literature that highlighted the connection between crime and drug use, and suggested the property crimes such as theft, burglary might be committed to raise funds to purchase drugs [32].



(a) Chicago (b) Cincinnati

Figure 4.10: Importance of the static features. Mean absolute values of input weights of FC layer in the static component.

The importance of the **static features** is explored by analyzing mean absolute input weights of fully connected layer (FC) in the static component (see Figure 4.10). For Chicago case, the demographic features (*Population, Gender Diversity* and *Race Diversity*) are significant. It is observe that *Owner Occupied H. units, Poverty* and *Educational Attainment* are also informative. For Cincinnati case, *Gender Div.* and *Population* are important features for forecasting as well as the *Educational Attainment* and *Per Capita Income*. Furthermore, additional analysis is performed for each individual static feature to see the relationship between the predictions made by the models and the static feature values. In order to do that, the neighborhoods are divided into five quantiles based on their static feature values, and the box-plots are provided in terms of overdose predictions for each quantile (see Figure C.1 and Figure C.2 in the Appendix). For Chicago case, the only feature showing a linear correlation with the predicted overdoses is population. As the population increases among the neighborhoods, the predicted number of overdose incidents increases. On the other hand, for Cincinnati case, population and owner occupied housing unites reveal a linear relationship with predicted number of overdoses. As the population increases or owner occupied housing units decreases, the predicted number of overdoses increases. Therefore, the neighborhoods with higher population, and lower owner occupied housing units may require additional resources to prevent opioid overdose. The relationship between rest of the static features and the predicted number of overdoses is non-linear. Moreover, the results shows that economic status is an important feature for neighborhoods of both cities, which is consistent with the previous work that suggested communities with a higher concentration of economic stressors (e.g. low income, poverty) may be vulnerable to abuse of opioids as a way to manage chronic stress and mood disorders [141]. Although there exist three economic status indicators, Group Lasso selects only one of them, *Poverty* for Chicago and *Per Capita Income* for Cincinnati.

4.5 Discussion and Conclusion

In this chapter, a community-attentive spatio-temporal predictive model, called CASTNet, was presented to forecast future societal events. The proposed model was applied to forecast opioid overdose events from crime dynamics as one of the application domains for societal events. The proposed model is a novel deep learning architecture based on multi-head attentional networks that learns different representation subspaces of features (communities) and allows the target locations to select location-specific community contributions for forecasting local incidents. At the same time, CASTNet allows for interpreting predictive features in both local-level and community-level, as well as the community memberships and the community contributions to forecasting local incidents. The extensive experiment results demonstrated the strength of CASTNet. It achieved superior forecasting performance on two real-world opioid overdose datasets compared to the several baseline and stat-of-the-art methods.

The experiment results suggested different spatio-temporal crime-overdose potential links. The overdose deaths at a target neighborhood in Chicago appeared to be better predicted by crime incidents at neighborhoods that share the same community with the target neighborhood. Also, change in crime incidents in those neighborhoods with low crime rates was an important indicator of future overdose deaths in most of the other neighborhoods. On the other hand, in Cincinnati, the crime incidents occurred in communities comprising those crime hotspots seemed to well predict the overdose events in most of the neighborhoods. Furthermore, the predictive local activities were different in two cases. While the local crime incidents, in particularly *Narcotics* and *Assault*, were predictive for local overdose deaths in Chicago, previous overdose occurrences were informative for future overdose incidents in Cincinnati. On the other hand, the global contributions to forecasting local overdose incidents showed similar patterns in both cities. Change in property crimes, in particular *Theft, Deceptive Practice, Burglary* and *Weapons Violation* (crime against to society) in Chicago, *Theft* and *Burglary* in Cincinnati, can be significant indicators for future local overdose incidents as well as certain type of violent crimes (*Battery* for Chicago and *Agg. Assault for Cincinnati*). Last but not the least, demographic characteristics, economic status and educational attainment of the neighborhoods in both cities may help forecasting the future local incidents. Findings from the experiments support the hypothesis that criminal activities and opioid overdose incidents may reveal spatio-temporal lags, and they are consistent with the literature.

4.5.1 Limitations

There are several limitations in the current work. (1) Although this work makes an in-depth comparison between two cases (Chicago vs. Cincinnati) in terms of potential spatio-temporal crime vs. opioid overdose links, there are differences in the data utilized for the different cases. There is no a systematic monitoring of drug abuse either in regional-level or local-level in the U.S. Also, there is no common reporting mechanism for incidents for different cities. Therefore, this study forecasts opioid overdose deaths for Chicago and heroin (a special type of opioids) overdoses for Cincinnati. In addition, although crime data is meticulously collected, organized and stored, there still exist different crime types for different city police departments. Therefore, in order to obtain more concrete insights and comparisons, there is a need for more systematic and compatible data collection mechanism across different locations. (2) The current proposed model does not consider the multi-resolution spatio-temporal dynamics for the prediction. Utilizing information from different granularity of time (e.g. day, week, year) and space (e.g neighborhood, city) may result in capturing and modeling better spatio-temporal dynamics. Although CASTNet can work with any type of spatial or temporal units (depending on the application domain), extending the current proposed architecture with multi-resolution setting may help to capture better spatio-temporal characteristics and increase the forecasting performance.

CHAPTER 5

CONCLUSION

In this thesis, novel interpretable spatio-temporal predictive deep learning models were proposed for learning the relationships between individual activities and societal events as well as forecasting the future societal events from these activities. The proposed models differentiated the local and the global feature contributions in spatio-temporal societal event forecasting domain for the first time. The proposed models were evaluated on the specific societal event domains (i.e. social movements and opioid overdoses) with multiple cases. In-depth analyses and comparisons were presented across the real-world cases.

First, ActAttn was proposed which (i) differentiates the local and the global feature contributions and (ii) identifies the hub locations through hierarchical attention mechanism. ActAttn could identify the contribution of each location to forecasting future events in other locations. ActAttn also incorporated Group Lasso regularization to select most informative intra-region and inter-region activity features as well as the location-specific static features. The proposed model was evaluated on the social movements domain with three real-world cases, where the future offline protests were predicted from the spatio-temporal social media activities and location-specific static information. Theory-driven feature extraction was employed to make sense of the association between the types of activity traces in social media and future offline protests. ActAttn yielded a boost in forecasting performance compared to the baseline methods. In terms of the activity features, LIWC-related features (e.g. anger, negative emotion, social) and Twitter engagement features (e.g. number of tweets) were the most informative feature groups in common across the different social movements, which is also consistent with the prior empirical studies [19]. Moreover, a few of the locations (states) were identified as the hubs that have more salient global contribution to forecasting future protests, and these locations differed in each case.

Second, CASTNet was proposed, inspired by the idea of multi-head attentional networks [121]. It aims to learn different representation subspaces of global dynamic features, where each subspace involves a set of locations called "community" that share similar behaviors. CASTNet allows the prediction for a given location to be individually optimized by the features contributed by a mixture of communities. To do so, it automatically differentiates the pairwise relationship between a particular event location and the other locations unlike the previous work [142, 64, 65]. Moreover, CASTNet allows for exploring community memberships, and community contributions as the global contributions to forecasting future local events. It also focuses on the most informative time steps when the local-level and the community-level activities are more predictive for the future opioid overdoses in the target locations. By leveraging Group Lasso regularization, the proposed model allows for interpreting predictive features in both local-level and community-level. The proposed model was evaluated on the opioid crisis domain with three real-world cases, where the future opioid overdose events were predicted from the spatio-temporal crime activities. CASTNet outperformed several baseline and stat-of-the-art methods on two real-world cases. Different spatio-temporal crime-overdose potential links were explored based on the experiment results. While local narcotics and assault crimes are very predictive for forecasting future overdoses in the same neighborhoods in Chicago, previous local overdose events are predictive in Cincinnati. Furthermore, overdose-related deaths at a target neighborhoods in Chicago could be be better predicted by crime incidents at neighborhoods that share the same community with the target neighborhood. In Cincinnati, the crimes committed in the crime hot-spots predicted the future opioid overdose events well in most of the target neighborhoods. Change in property crimes in other locations are important indicators for future local overdose events for both cities. In relation to static features, demographic characteristics, economic status and educational attainment seemed informative for forecasting the future local events in both cities. Also, the experiment results are consistent with the literature.

5.1 Limitations and Future Work

Despite the notable contributions of the proposed models in this thesis, there exist several limitations of them. These limitations and the possible relevant future directions can be described as follows:

Each proposed model achieved significant performance increase, and provided meaningful insights about how events unfolded over time and across space on a specific societal event domain with multiple cases. Although the proposed models can be applied on any type of societal event domain, a further evaluation is needed on various societal event domains in order to validate their efficiency and generalizability. A possible future direction of this study can be application and evaluation of the suggested methods on the different societal event domains, and even other types of event domains which have the spatio-temporal structure by considering the potential links between the activities and the relevant events.

Another limitation is that the proposed models performed short-term event forecasting tasks within the scope this thesis. Longer-term event forecasting performances of the proposed models are not examined in this work due to the limitation in data, in particularly social movement data. However, to make more efficient predictions, longer-term forecasting may be required depending on the application domain. A potential future work would be the evaluation of the suggested models on the longer-term prediction tasks, and enhancing the current architectures of the models with the components or the mechanisms (e.g. incorporating decoder networks into the final layers) for this task in order to achieve accurate and efficient longer-term predictions.

Another limitation of the suggested models is using single-resolution spatio-temporal data for modeling and forecasting. The proposed models can work with any types of temporal and spatial units depending on the application domain. However, the current proposed models do not consider the multi-resolution spatio-temporal dynamics for the prediction. Utilizing information from different granularity of time (e.g. day, week, year) and space (e.g neighborhood, city) may result in capturing and modeling better spatio-temporal dynamics. Therefore, it may increase the forecasting performance of the suggested models. In this manner, another future research direction would be enhancing the current proposed architectures with the multi-resolution setting so that they may help to capture better spatio-temporal characteristics and increase the forecasting performance.

Furthermore, selection of the hyper-parameters and resolutions in the experiments is limited. First, the suggested models used fixed hyper-parameters, which were tuned based on the development sets. However, hyper-parameters could be adaptively updated with respect to more recent data points, which may lead to an increase in the forecasting performance due to temporally evolving nature of the data. Therefore, a possible future work would be the investigation of the effect of adaptive selection of hyper-parameters on the forecasting performance. Second, the proposed models were evaluated utilizing specific resolutions in time (e.g. day, week) and space (e.g. neighborhood, state) with a small set of values of the hyper-parameters including lead time and window size. These models outperformed the other methods in the literature with the corresponding settings. However, as a future work, further evaluations and analyses are needed in order to observe how the proposed models would perform with different resolution settings from a more comprehensive set of hyper-parameter values.

Missing information is a common challenge in predicting/forecasting societal events. The suggested models were not developed explicitly considering the information missingness such as *missing at random* and *missing not at random*. However, the proposed architectures model the complex interactions between temporal and spatial dimensions as well as the relationships between local and global activities over time and across space. Therefore, these models are expected to be robust to the missing information. It is also shown that ActAttn is useful in dealing with the missing information, and is more robust compared to the other methods. A further research would be the improvement of the suggested models with components specifically targeting the missingness problems. Lastly, the suggested models were not evaluated with regard to cold start problem. In other words, all predictions were performed for the locations which have samples in the training sets. In the experiments, it is observed that global (inter-region) contributions play a significant role on forecasting events in the target locations. Therefore, the proposed models would result in successful predictions at some degree through the global components for the novel locations, which are not seen in the training sets. Yet, a possible future work would be evaluation and analysis of the suggested models in terms of cold start problem.

REFERENCES

[1] A. Doyle, G. Katz, K. Summers, C. Ackermann, I. Zavorin, Z. Lim, S. Muthiah, P. Butler, N. Self, L. Zhao, *et al.*, "Forecasting significant societal events using the embers streaming predictive analytics system," *Big data*, vol. 2, no. 4, pp. 185–195, 2014.

[2] Y. Ning, S. Muthiah, H. Rangwala, and N. Ramakrishnan, "Modeling precursors for event forecasting via nested multi-instance learning," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1095–1104, ACM, 2016.

[3] Wikipedia, "Ferguson unrest — Wikipedia, the free encyclopedia." `http://en.wikipedia.org/w/index.php?title=Ferguson%20unrest&oldid=919310025`, 2018. [Online; accessed 01-April-2018].

[4] T. N. Winkelman, V. W. Chang, and I. A. Binswanger, "Health, polysubstance use, and criminal justice involvement among adults with varying levels of opioid use," *JAMA network open*, vol. 1, no. 3, pp. e180558–e180558, 2018.

[5] D. S. Burke, "Forecasting the opioid epidemic," *Science*, vol. 354, no. 6312, pp. 529–529, 2016.

[6] M. S. Lewis-Beck, C. Tien, and D. Pervin, "Election forecasting: The long-view," in *Oxford Handbooks Online*, Oxford University Press Oxford, 2016.

[7] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan, "Spatiotemporal event forecasting in social media," in *Proceedings of the 2015 SIAM International Conference on Data Mining*, pp. 963–971, SIAM, 2015.

[8] 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*, pp. 1503–1512, ACM, 2015.

[9] J. Zhang, Y. Zheng, D. Qi, R. Li, and X. Yi, "Dnn-based prediction model for spatio-temporal data," in *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, p. 92, ACM, 2016.

[10] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan, "Multi-resolution spatial event forecasting in social media," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 689–698, IEEE, 2016.

[11] D. A. Snow, S. A. Soule, and H. Kriesi, *The Blackwell companion to social movements*. John Wiley & Sons, 2008.

[12] S. Valenzuela, "Unpacking the use of social media for protest behavior: The roles of information, opinion expression, and activism," *American Behavioral Scientist*, vol. 57, no. 7, pp. 920–942, 2013.

[13] Y. Theocharis, W. Lowe, J. W. Van Deth, and G. García-Albacete, “Using twitter to mobilize protest action: online mobilization patterns and action repertoires in the occupy wall street, indignados, and aganaktismenoi movements,” *Information, Communication & Society*, vol. 18, no. 2, pp. 202–220, 2015.

[14] J. Van Stekelenburg and B. Klandermans, “The social psychology of protest,” *Current Sociology*, vol. 61, no. 5-6, pp. 886–905, 2013.

[15] J. Van Stekelenburg, “The political psychology of protest,” *European Psychologist*, 2013.

[16] J. Goodwin and J. M. Jasper, “Emotions and social movements,” in *Handbook of the Sociology of Emotions*, pp. 611–635, Springer, 2006.

[17] M. D. Conover, E. Ferrara, F. Menczer, and A. Flammini, “The digital evolution of occupy wall street,” *PloS One*, vol. 8, no. 5, p. e64679, 2013.

[18] W.-T. Chung, Y.-R. Lin, A. Li, A. M. Ertugrul, and M. Yan, “March with and without feet: the talking about protests and beyond,” in *International Conference on Social Informatics*, pp. 134–150, Springer, 2018.

[19] M. De Choudhury, S. Jhaver, B. Sugar, and I. Weber, “Social media participation in an activist movement for racial equality,” in *Tenth International AAAI Conference on Web and Social Media*, 2016.

[20] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno, “The dynamics of protest recruitment through an online network,” *Scientific Reports*, vol. 1, p. 197, 2011.

[21] M. D. Conover, C. Davis, E. Ferrara, K. McKelvey, F. Menczer, and A. Flammini, “The geospatial characteristics of a social movement communication network,” *PloS One*, vol. 8, no. 3, p. e55957, 2013.

[22] J. He, L. Hong, V. Frias-Martinez, and P. Torrens, “Uncovering social media reaction pattern to protest events: a spatiotemporal dynamics perspective of ferguson unrest,” in *International Conference on Social Informatics*, pp. 67–81, Springer, 2015.

[23] H. Qi, P. Manrique, D. Johnson, E. Restrepo, and N. F. Johnson, “Open source data reveals connection between online and on-street protest activity,” *EPJ Data Science*, vol. 5, no. 1, p. 18, 2016.

[24] M. Warner, L. H. Chen, D. M. Makuc, R. N. Anderson, and A. M. Miniño, “Drug poisoning deaths in the united states, 1980-2008.,” *NCHS Data Brief*, no. 81, pp. 1–8, 2011.

[25] R. A. Rudd, N. Aleshire, J. E. Zibbell, and R. Matthew Gladden, “Increases in drug and opioid overdose deaths—united states, 2000–2014,” *American Journal of Transplantation*, vol. 16, no. 4, pp. 1323–1327, 2016.

[26] W. M. Compton, C. M. Jones, and G. T. Baldwin, “Relationship between nonmedical prescription-opioid use and heroin use,” *New England Journal of Medicine*, vol. 374, no. 2, pp. 154–163, 2016.

[27] H. Hedegaard, L.-H. Chen, and M. Warner, “Drug-poisoning deaths involving heroin: United states, 2000–2013,” 2015.

[28] Centers for Disease Control and Prevention, “Web-based Injury Statistics Query and Reporting System (WISQARS).” `http://www.cdc.gov/injury/wisqars/fatal.html`. Accessed: 2018-05-01.

[29] A. Kolodny, D. T. Courtwright, C. S. Hwang, P. Kreiner, J. L. Eadie, T. W. Clark, and G. C. Alexander, “The prescription opioid and heroin crisis: a public health approach to an epidemic of addiction,” *Annual Review of Public Health*, vol. 36, pp. 559–574, 2015.

[30] H. Jalal, J. M. Buchanich, M. S. Roberts, L. C. Balmert, K. Zhang, and D. S. Burke, “Changing dynamics of the drug overdose epidemic in the united states from 1979 through 2016,” *Science*, vol. 361, no. 6408, p. eaau1184, 2018.

[31] R. Hammersley, A. Forsyth, V. Morrison, and J. B. Davies, “The relationship between crime and opioid use,” *Addiction*, vol. 84, no. 9, pp. 1029–1043, 1989.

[32] T. Bennett, K. Holloway, and D. Farrington, “The statistical association between drug misuse and crime: A meta-analysis,” *Aggression and Violent Behavior*, vol. 13, no. 2, pp. 107–118, 2008.

[33] M. Pierce, K. Hayhurst, S. M. Bird, M. Hickman, T. Seddon, G. Dunn, and T. Millar, “Quantifying crime associated with drug use among a large cohort of sanctioned offenders in england and wales,” *Drug and Alcohol Dependence*, vol. 155, pp. 52–59, 2015.

[34] T. Seddon, “Drugs, crime and social exclusion: social context and social theory in british drugs–crime research,” *British Journal of Criminology*, vol. 46, no. 4, pp. 680–703, 2005.

[35] J. M. Glanz, K. J. Narwaney, S. R. Mueller, E. M. Gardner, S. L. Calcaterra, S. Xu, K. Breslin, and I. A. Binswanger, “Prediction model for two-year risk of opioid overdose among patients prescribed chronic opioid therapy,” *Journal of General Internal Medicine*, vol. 33, no. 10, pp. 1646–1653, 2018.

[36] P. J. Gruenewald, “Geospatial analyses of alcohol and drug problems: empirical needs and theoretical foundations,” *GeoJournal*, vol. 78, no. 3, pp. 443–450, 2013.

[37] P. J. Gruenewald, W. R. Ponicki, L. G. Remer, L. A. Waller, L. Zhu, and D. M. Gorman, “Mapping the spread of methamphetamine abuse in california from 1995 to 2008,” *American Journal of Public Health*, vol. 103, no. 7, pp. 1262–1270, 2013.

[38] H. Choi, K. Cho, and Y. Bengio, “Fine-grained attention mechanism for neural machine translation,” *Neurocomputing*, vol. 284, pp. 171–176, 2018.

[39] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal, “Semeval-2019 task 3: Emocontext contextual emotion detection in text,” in *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 39–48, 2019.

[40] W.-S. Chu, F. De la Torre, and J. F. Cohn, “Learning spatial and temporal cues for multi-label facial action unit detection,” in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pp. 25–32, IEEE, 2017.

[41] M. Shan and N. Atanasov, “A spatiotemporal model with visual attention for video classification,” *arXiv preprint arXiv:1707.02069*, 2017.

[42] L. Hu, J. Li, L. Nie, X.-L. Li, and C. Shao, “What happens next? future subevent prediction using contextual hierarchical lstm,” in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[43] M. Granroth-Wilding and S. Clark, “What happens next? event prediction using a compositional neural network model,” in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[44] N. Du, H. Dai, R. Trivedi, U. Upadhyay, M. Gomez-Rodriguez, and L. Song, “Recurrent marked temporal point processes: Embedding event history to vector,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1555–1564, ACM, 2016.

[45] S. Xiao, J. Yan, M. Farajtabar, L. Song, X. Yang, and H. Zha, “Learning time series associated event sequences with recurrent point process networks,” *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[46] Y. Gao, L. Zhao, L. Wu, Y. Ye, H. Xiong, and C. Yang, “gao2019incomplete,” in *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[47] B. Wang, D. Zhang, D. Zhang, P. J. Brantingham, and A. L. Bertozzi, “Deep learning for real time crime forecasting,” *arXiv preprint arXiv:1707.03340*, 2017.

[48] C. Huang, C. Zhang, J. Zhao, X. Wu, N. Chawla, and D. Yin, “Mist: A multiview and multimodal spatial-temporal learning framework for citywide abnormal event forecasting,” in *The World Wide Web Conference*, pp. 717–728, ACM, 2019.

[49] C. Huang, J. Zhang, Y. Zheng, and N. V. Chawla, “Deepcrime: attentive hierarchical recurrent networks for crime prediction,” in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1423–1432, ACM, 2018.

[50] H.-W. Kang and H.-B. Kang, “Prediction of crime occurrence from multi-modal data using deep learning,” *PloS One*, vol. 12, no. 4, p. e0176244, 2017.

[51] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, “Structural-rnn: Deep learning on spatio-temporal graphs,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5308–5317, 2016.

[52] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, “Deep learning architecture for air quality predictions,” *Environmental Science and Pollution Research*, vol. 23, no. 22, pp. 22408–22417, 2016.

[53] A. Ghaderi, B. M. Sanandaji, and F. Ghaderi, “Deep forecast: Deep learning-based spatio-temporal forecasting,” *arXiv preprint arXiv:1707.08110*, 2017.

[54] H. Yu, Z. Wu, S. Wang, Y. Wang, and X. Ma, “Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks,” *Sensors*, vol. 17, no. 7, p. 1501, 2017.

[55] M. Das and S. K. Ghosh, “Deep-step: A deep learning approach for spatiotemporal prediction of remote sensing data,” *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1984–1988, 2016.

[56] J. Fan, Q. Li, J. Hou, X. Feng, H. Karimian, and S. Lin, “A spatiotemporal prediction framework for air pollution based on deep rnn,” *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 4, p. 15, 2017.

[57] T.-C. Bui, V.-D. Le, and S.-K. Cha, “A deep learning approach for forecasting air pollution in south korea using lstm,” *arXiv preprint arXiv:1804.07891*, 2018.

[58] J. Wang, J. Tang, Z. Xu, Y. Wang, G. Xue, X. Zhang, and D. Yang, “Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach,” in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pp. 1–9, IEEE, 2017.

[59] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, “Traffic flow prediction with big data: a deep learning approach,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2014.

[60] J. Zhang, Y. Zheng, and D. Qi, “Deep spatio-temporal residual networks for citywide crowd flows prediction,” in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[61] X. Zhou, Y. Shen, Y. Zhu, and L. Huang, “Predicting multi-step citywide passenger demands using attention-based neural networks,” in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 736–744, ACM, 2018.

[62] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, “Convolutional lstm network: A machine learning approach for precipitation nowcasting,” in *Advances in Neural Information Processing Systems*, pp. 802–810, 2015.

[63] N. Laptev, J. Yosinski, L. E. Li, and S. Smyl, “Time-series extreme event forecasting with neural networks at uber,” in *International Conference on Machine Learning*, vol. 34, pp. 1–5, 2017.

[64] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, “A dual-stage attention-based recurrent neural network for time series prediction,” in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017.

[65] Y. Liang, S. Ke, J. Zhang, X. Yi, and Y. Zheng, “Geoman: multi-level attention networks for geo-sensory time series prediction,” in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 3428–3434, AAAI Press, 2018.

[66] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007.

[67] A. G. Hawkes, “Spectra of some self-exciting and mutually exciting point processes,” *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[68] Y. Ogata, “Space-time point-process models for earthquake occurrences,” *Annals of the Institute of Statistical Mathematics*, vol. 50, no. 2, pp. 379–402, 1998.

[69] M. Short, G. Mohler, P. J. Brantingham, and G. Tita, “Gang rivalry dynamics via coupled point process networks.,” *Discrete & Continuous Dynamical Systems-Series B*, vol. 19, no. 5, 2014.

[70] E. Bacry, T. Jaisson, and J.-F. Muzy, “Estimation of slowly decreasing hawkes kernels: application to high-frequency order book dynamics,” *Quantitative Finance*, vol. 16, no. 8, pp. 1179–1201, 2016.

[71] F. Musmeci and D. Vere-Jones, “A space-time clustering model for historical earthquakes,” *Annals of the Institute of Statistical Mathematics*, vol. 44, no. 1, pp. 1–11, 1992.

[72] E. Lewis, G. Mohler, P. J. Brantingham, and A. L. Bertozzi, “Self-exciting point process models of civilian deaths in iraq,” *Security Journal*, vol. 25, no. 3, pp. 244–264, 2012.

[73] B. B. Batu, T. T. Temizel, and H. Ş. Düzgün, “A non-parametric algorithm for discovering triggering patterns of spatio-temporal event types,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2629–2642, 2017.

[74] J. F. C. Kingman, *Poisson processes*, vol. 3. Clarendon Press, 1992.

[75] A. Ferraz Costa, Y. Yamaguchi, A. Juci Machado Traina, C. Traina Jr, and C. Faloutsos, “Rsc: Mining and modeling temporal activity in social media,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 269–278, ACM, 2015.

[76] R. F. Engle and J. R. Russell, “Autoregressive conditional duration: a new model for irregularly spaced transaction data,” *Econometrica*, pp. 1127–1162, 1998.

[77] S. Xiao, J. Yan, X. Yang, H. Zha, and S. M. Chu, “Modeling the intensity function of point process via recurrent neural networks.,” in *AAAI*, pp. 1597–1603, 2017.

[78] F. Qiao, P. Li, X. Zhang, Z. Ding, J. Cheng, and H. Wang, “Predicting social unrest events with hidden markov models using gdelt,” *Discrete Dynamics in Nature and Society*, vol. 2017, 2017.

[79] E. Alevizos, A. Artikis, and G. Paliouras, “Event forecasting with pattern markov chains,” in *Proceedings of the 11th ACM International Conference on Distributed and Event-based Systems*, pp. 146–157, ACM, 2017.

[80] M. J. Sanjari and H. Gooi, “Probabilistic forecast of pv power generation based on higher order markov chain,” *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 2942–2952, 2016.

[81] L. Yang, M. He, J. Zhang, and V. Vittal, “Support-vector-machine-enhanced markov model for short-term wind power forecast,” *IEEE Transactions on Sustainable Energy*, vol. 6, no. 3, pp. 791–799, 2015.

[82] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme, “Factorizing personalized markov chains for next-basket recommendation,” in *Proceedings of the 19th International Conference on World Wide Web*, pp. 811–820, ACM, 2010.

[83] S. Sahebi, Y.-R. Lin, and P. Brusilovsky, “Tensor factorization for student modeling and performance prediction in unstructured domain.,” *International Educational Data Mining Society*, 2016.

[84] M. T. Bahadori, Q. R. Yu, and Y. Liu, “Fast multivariate spatio-temporal analysis via low rank tensor learning,” in *Advances in Neural Information Processing Systems*, pp. 3491–3499, 2014.

[85] V. W. Zheng, B. Cao, Y. Zheng, X. Xie, and Q. Yang, “Collaborative filtering meets mobile recommendation: A user-centered approach.,” in *AAAI*, vol. 10, pp. 236–241, 2010.

[86] J. Xu, J. Zhou, P.-N. Tan, X. Liu, and L. Luo, “Wisdom: Weighted incremental spatio-temporal multi-task learning via tensor decomposition,” in *2016 IEEE International Conference on Big Data (Big Data)*, pp. 522–531, IEEE, 2016.

[87] J. Xu, X. Liu, T. Wilson, P.-N. Tan, P. Hatami, and L. Luo, “Muscat: Multi-scale spatio-temporal learning with application to climate modeling.,” in *IJCAI*, pp. 2912–2918, 2018.

[88] Q. Liu, S. Wu, L. Wang, and T. Tan, “Predicting the next location: A recurrent model with spatial and temporal contexts.,” in *AAAI*, pp. 194–200, 2016.

[89] M. Arias, A. Arratia, and R. Xuriguera, “Forecasting with twitter data,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 1, p. 8, 2013.

[90] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[91] S. Aghababaei and M. Makrehchi, “Mining social media content for crime prediction,” in *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 526–531, IEEE, 2016.

[92] X. Wang, M. S. Gerber, and D. E. Brown, “Automatic crime prediction using events extracted from twitter posts,” in *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pp. 231–238, Springer, 2012.

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

[94] G. Korkmaz, J. Cadena, C. J. Kuhlman, A. Marathe, A. Vullikanti, and N. Ramakrishnan, “Combining heterogeneous data sources for civil unrest forecasting,” in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pp. 258–265, ACM, 2015.

[95] R. Korolov, D. Lu, J. Wang, G. Zhou, C. Bonial, C. Voss, L. Kaplan, W. Wallace, J. Han, and H. Ji, “On predicting social unrest using social media,” in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 89–95, IEEE, 2016.

[96] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz, *et al.*, “’beating the news’ with embers: forecasting civil unrest using open source indicators,” in *Proceedings of the 20th*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1799–1808, 2014.

[97] J. Cadena, G. Korkmaz, C. J. Kuhlman, A. Marathe, N. Ramakrishnan, and A. Vullikanti, “Forecasting social unrest using activity cascades,” *PloS One*, vol. 10, no. 6, p. e0128879, 2015.

[98] T. Rekatsinas, S. Ghosh, S. R. Mekaru, E. O. Nsoesie, J. S. Brownstein, L. Getoor, and N. Ramakrishnan, “Sourceseer: Forecasting rare disease outbreaks using multiple data sources,” in *Proceedings of the 2015 SIAM International Conference on Data Mining*, pp. 379–387, SIAM, 2015.

[99] L. Zhao, J. Wang, F. Chen, C.-T. Lu, and N. Ramakrishnan, “Spatial event forecasting in social media with geographically hierarchical regularization,” *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1953–1970, 2017.

[100] L. Zhao, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan, “Hierarchical incomplete multi-source feature learning for spatiotemporal event forecasting,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2085–2094, ACM, 2016.

[101] L. Zhao, Q. Sun, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan, “Feature constrained multi-task learning models for spatiotemporal event forecasting,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 5, pp. 1059–1072, 2017.

[102] Q. Zhang, N. Perra, D. Perrotta, M. Tizzoni, D. Paolotti, and A. Vespignani, “Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model,” in *Proceedings of the 26th International Conference on World Wide Web*, pp. 311–319, International World Wide Web Conferences Steering Committee, 2017.

[103] X. Zhang, L. Zhao, A. P. Boedihardjo, C.-T. Lu, and N. Ramakrishnan, “Spatiotemporal event forecasting from incomplete hyper-local price data,” in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 507–516, ACM, 2017.

[104] L. Zhao, J. Wang, and X. Guo, “Distant-supervision of heterogeneous multitask learning for social event forecasting with multilingual indicators,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[105] K. Cho, “Natural language understanding with distributed representation,” *arXiv preprint arXiv:1511.07916*, 2015.

[106] J. L. Elman, “Finding structure in time,” *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.

[107] J. S. Bridle, “Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition,” in *Neurocomputing*, pp. 227–236, Springer, 1990.

[108] M. C. Mozer, “A focused backpropagation algorithm for temporal,” *Backpropagation: Theory, Architectures, and Applications*, vol. 137, 1995.

[109] P. J. Werbos, "Generalization of backpropagation with application to a recurrent gas market model," *Neural Networks*, vol. 1, no. 4, pp. 339–356, 1988.

[110] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural Networks: Tricks of the Trade*, pp. 9–48, Springer, 2012.

[111] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[112] Y. Bengio, P. Simard, P. Frasconi, *et al.*, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[113] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks.," in *IJCAI*, pp. 3818–3824, 2016.

[114] W. Hu, K. K. Singh, F. Xiao, J. Han, C.-N. Chuah, and Y. J. Lee, "Who will share my image?: Predicting the content diffusion path in online social networks," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 252–260, ACM, 2018.

[115] L. Gao, X. Wang, J. Song, and Y. Liu, "Fused gru with semantic-temporal attention for video captioning," *Neurocomputing*, 2019.

[116] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[117] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[118] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489, 2016.

[119] P. Rodríguez, J. M. Gonfaus, G. Cucurull, F. XavierRoca, and J. Gonzalez, "Attend and rectify: a gated attention mechanism for fine-grained recovery," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 349–364, 2018.

[120] O. Firat, K. Cho, and Y. Bengio, "Multi-way, multilingual neural machine translation with a shared attention mechanism," *arXiv preprint arXiv:1601.01073*, 2016.

[121] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[122] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[123] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[124] S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini, “Group sparse regularization for deep neural networks,” *Neurocomputing*, vol. 241, pp. 81–89, 2017.

[125] L. Zhao, Q. Hu, and W. Wang, “Heterogeneous feature selection with multi-modal deep neural networks and sparse group lasso,” *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1936–1948, 2015.

[126] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, “Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks,” in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[127] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[128] A. M. Ertugrul, Y.-R. Lin, W.-T. Chung, M. Yan, and A. Li, “Activism via attention: interpretable spatiotemporal learning to forecast protest activities,” *EPJ Data Science*, vol. 8, no. 1, p. 5, 2019.

[129] M. Denil, L. Bazzani, H. Larochelle, and N. de Freitas, “Learning where to attend with deep architectures for image tracking,” *Neural Computation*, vol. 24, no. 8, pp. 2151–2184, 2012.

[130] Wikipedia, “Unite the Right rally — Wikipedia, the free encyclopedia.” `http://en.wikipedia.org/w/index.php?title=Unite%20the%20Right%20rally&oldid=919735859`, 2018. [Online; accessed 01-April-2018].

[131] D. Freelon, C. D. McIlwain, and M. D. Clark, “Beyond the hashtags:# ferguson,# blacklivesmatter, and the online struggle for offline justice,” *Center for Media & Social Impact, American University, Forthcoming*, 2016.

[132] C. Chung and J. W. Pennebaker, “The psychological functions of function words,” *Social Communication*, pp. 343–359, 2007.

[133] J. C. de Albornoz, L. Plaza, and P. Gervás, “Sentisense: An easily scalable concept-based affective lexicon for sentiment analysis.,” in *LREC*, pp. 3562–3567, 2012.

[134] J. Graham, J. Haidt, and B. A. Nosek, “Liberals and conservatives rely on different sets of moral foundations.,” *Journal of Personality and Social Psychology*, vol. 96, no. 5, p. 1029, 2009.

[135] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[136] A. M. Ertugrul, Y.-R. Lin, and T. Taskaya-Temizel, “Castnet: Community-attentive spatio-temporal networks for opioid overdose forecasting,” in *Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2019)*, 2019.

[137] A. Tuor, S. Kaplan, B. Hutchinson, N. Nichols, and S. Robinson, “Predicting user roles from computer logs using recurrent neural networks.,” in *AAAI*, pp. 4993–4994, 2017.

[138] T. Ochiai, S. Matsuda, H. Watanabe, and S. Katagiri, "Automatic node selection for deep neural networks using group lasso regularization," in *ICASSP*, 2017.

[139] G. E. Box and D. A. Pierce, "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models," *Journal of the American Statistical Association*, vol. 65, no. 332, pp. 1509–1526, 1970.

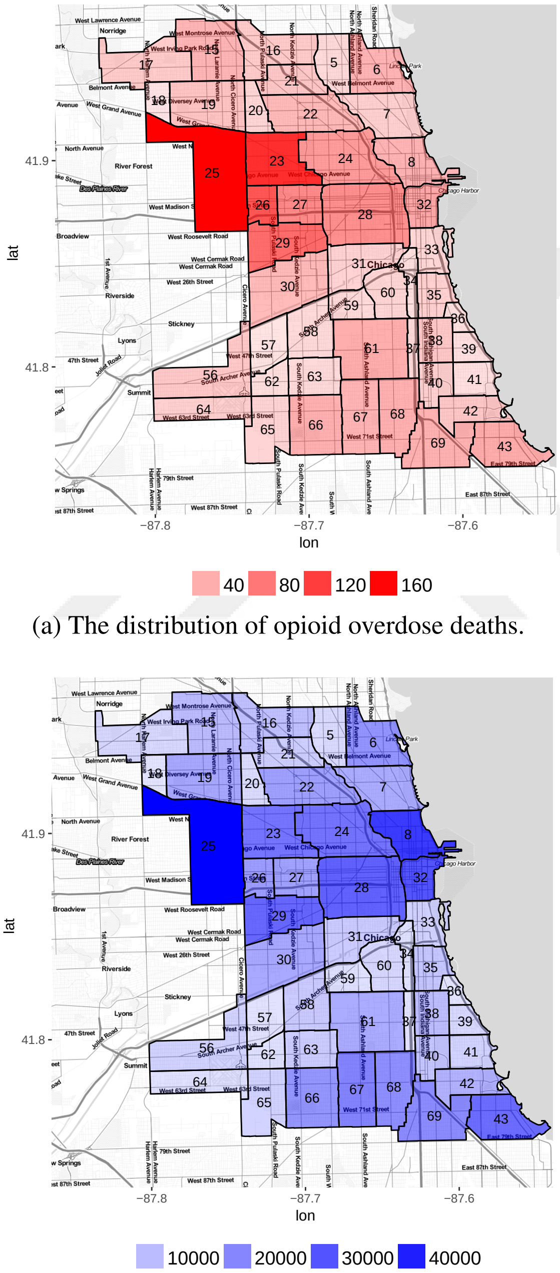
[140] E. Zivot and J. Wang, "Vector autoregressive models for multivariate time series," *Modeling Financial Time Series with S-Plus®*, pp. 385–429, 2006.

[141] N. B. King, V. Fraser, C. Boikos, R. Richardson, and S. Harper, "Determinants of increased opioid-related mortality in the united states and canada, 1990–2013: a systematic review," *American Journal of Public Health*, vol. 104, no. 8, pp. e32–e42, 2014.

[142] A. M. Ertugrul, Y.-R. Lin, C. Mair, and T. Taskaya Temizel, "Forecasting heroin overdose occurrences from crime incidents," in *SBP-BRiMS*, 2018.

Appendix A

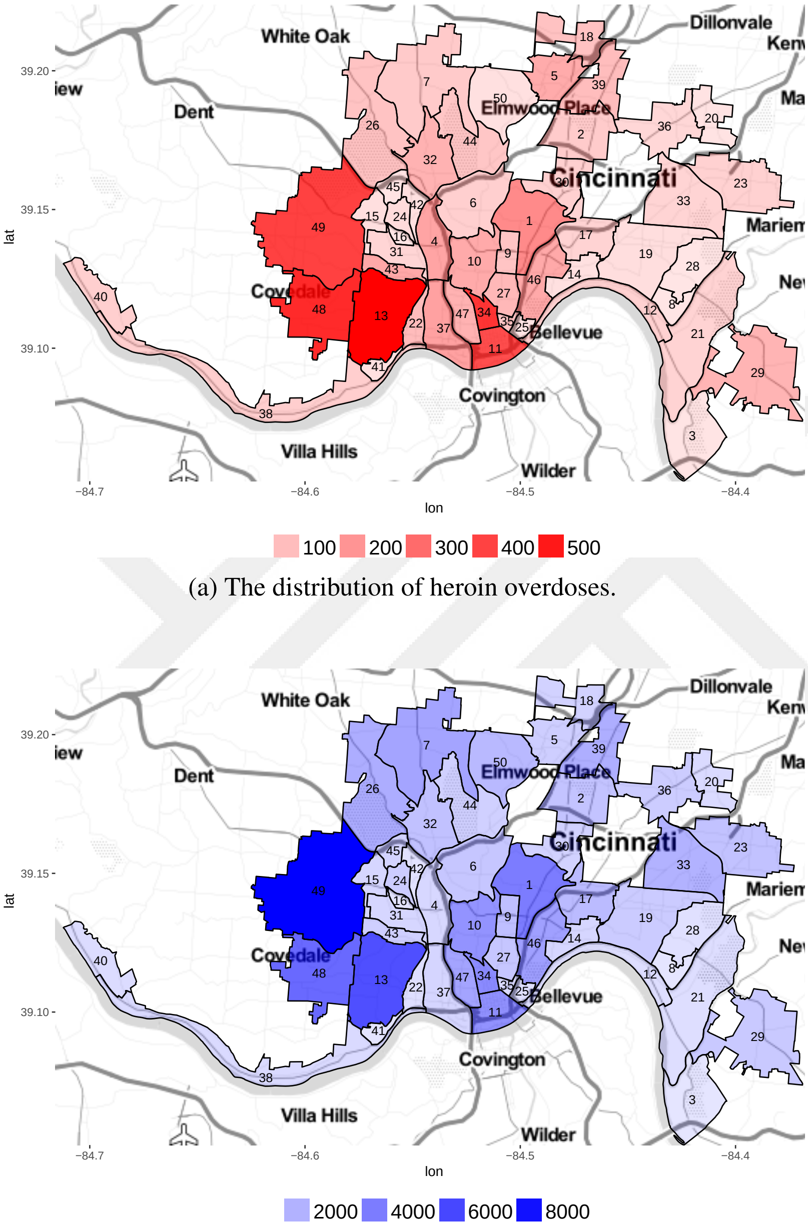
OPIOID OVERDOSE AND CRIME DISTRIBUTIONS



(a) The distribution of opioid overdose deaths.

(b) The distribution of crime incidents.

Figure A.1: The distribution of total number of (a) overdose deaths and (b) crime incidents in City of Chicago by neighborhoods. Rectangles in legends represent the amount of incidents.



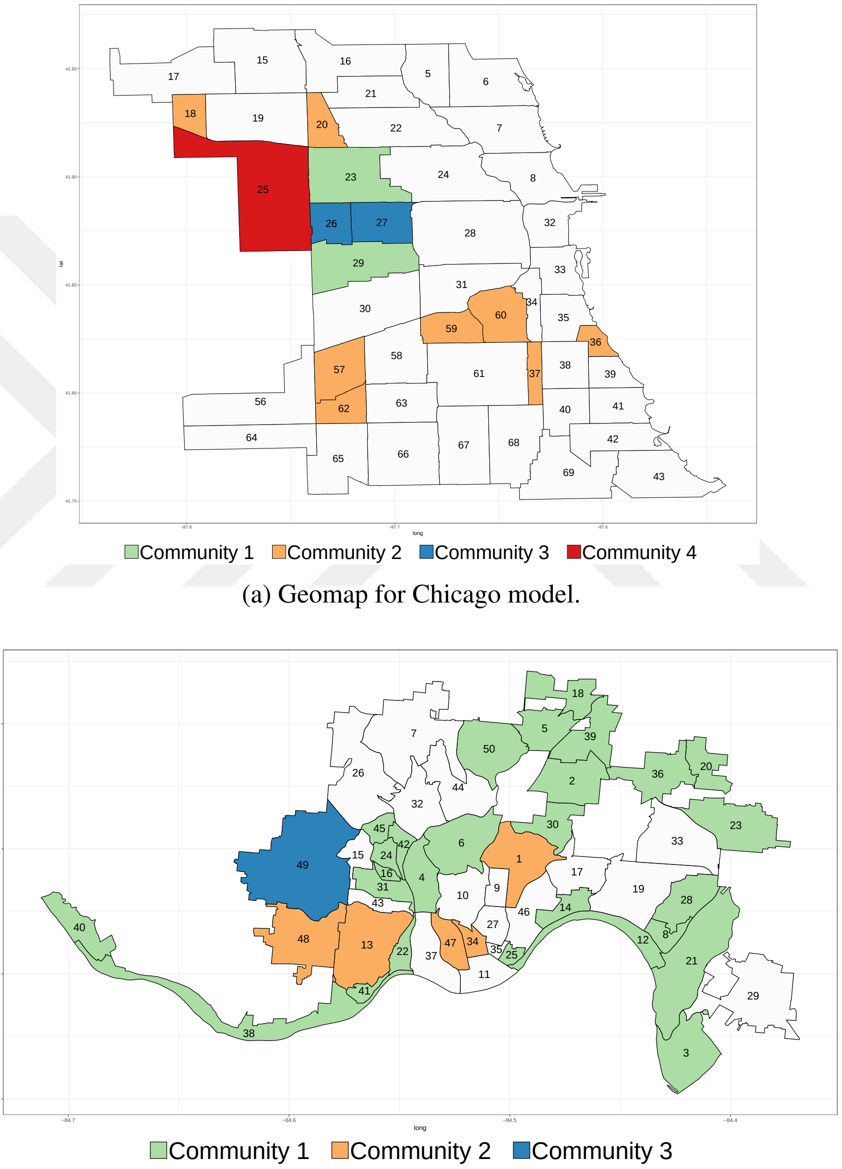
(a) The distribution of heroin overdoses.

(b) The distribution of crime incidents.

Figure A.2: The distribution of total number of (a) heroin overdoses and (b) crime incidents in City of Cincinnati by neighborhoods. Rectangles in the legends represent the amount of incidents.

Appendix B

GEOMAPS SHOWING COMMUNITY MEMBERSHIPS


🟩Community 1 🟧Community 2 🟦Community 3 🟥Community 4
long
lat
41.95
41.90
41.85
41.80
41.75
-87.8
-87.7
-87.6
🟩Community 1 🟧Community 2 🟦Community 3
39.20
39.15
39.10
39.05
-84.7
-84.6
-84.5
-84.4

(a) Geomap for Chicago model.

(b) Geomap for Cincinnati model.

Figure B.1: Geomaps showing the community memberships of the neighborhoods. Each neighborhood is colored by the community for which it has the highest membership among the other communities.

Appendix C

STATIC FEATURE ANALYSIS

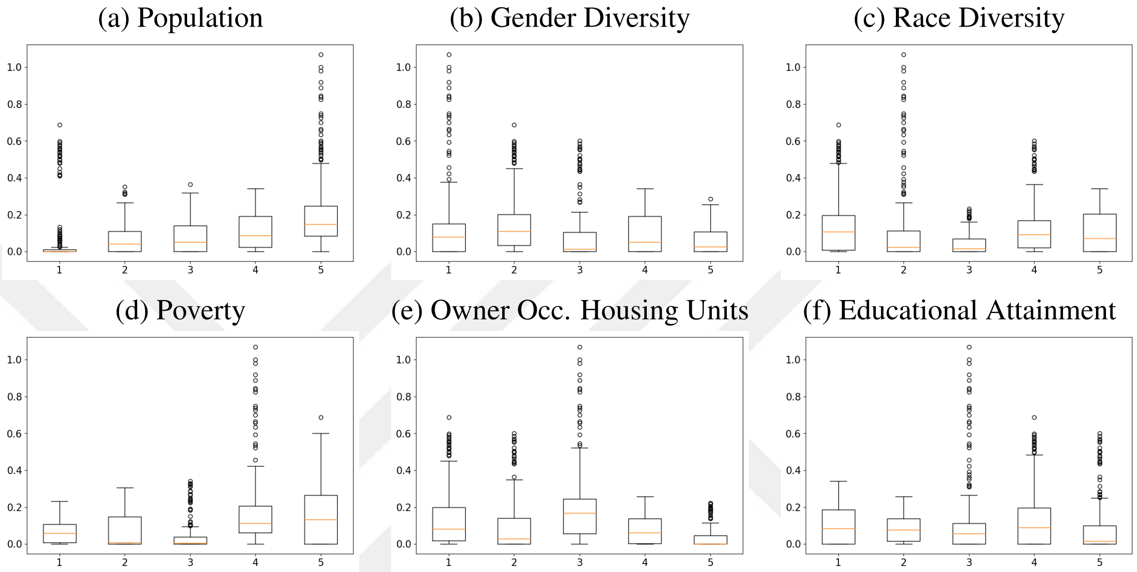
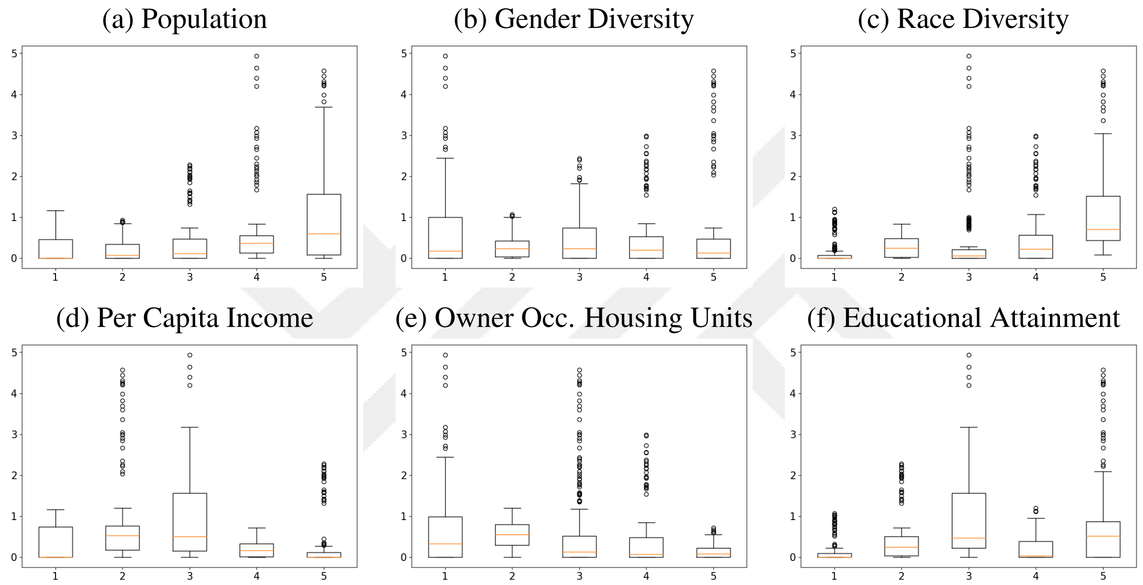


Figure C.1: Static feature analysis for the Chicago model. While x axis represents the quantiles for neighborhoods based on their static feature values, y axis indicates the number of opioid overdoses predicted by the model.


(a) Population

(b) Gender Diversity

(c) Race Diversity

(d) Per Capita Income

(e) Owner Occ. Housing Units

(f) Educational Attainment

Figure C.2: Static feature analysis for the Cincinnati model. While x axis represents the quantiles for neighborhoods based on their static feature values, y axis indicates the number of opioid overdoses predicted by the model.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Ertuğrul, Ali Mert
Nationality: Turkish (TC)
Date and Place of Birth: 30.06.1988, Konya
Marital Status: Married

EDUCATION

Degree	Institution	Year of Graduation
Ph.D.	Dept. of Information Systems, METU, Turkey	ongoing
M.S.	Dept. of Information Systems, METU, Turkey	2015
B.S.	Dept. of Computer Engineering, METU, Turkey	2011
High School	Konya Meram Science High School, Turkey	2005

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
Sep 2017 - ongoing	University of Pittsburgh, USA	Visiting Scholar
Dec 2013 - Oct 2017	Dept. of Information Systems, METU, Turkey	Research Assistant
Jul 2016 - Sep 2016	TU Delft, Netherlands	Visiting Ph.D. Student
Nov 2014 - Dec 2015	Bilgi Group Ltd, METU Technopolice, Turkey	Researcher & Analyst
Oct 2011 - Sep 2013	Arcelik A.S. Turkey	Software Engineer

PUBLICATIONS

1. **Ali Mert Ertugrul**, Yu-Ru Lin, Tugba Taskaya-Temizel. CASTNet: Community-Attentive Spatio-Temporal Networks for Opioid Overdose Forecasting, *17th Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, 2019.

2. **Ali Mert Ertugrul**, Yu-Ru Lin, Wen-Ting Chung, Muheng Yan, Ang Li. Activism via attention: interpretable spatiotemporal learning to forecast protest activities, *EPJ Data Science 8 (1), 5*, 2019.

3. **Ali Mert Ertugrul**, Yu-Ru Lin, Tugba Taskaya-Temizel. Forecasting Heroin Overdose Occurrences from Crime Incidents, *11th International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS)*, 2018.

4. Xian Teng, Muheng Yan, **Ali Mert Ertugrul**, Yu-Ru Lin. Deep into Hypersphere: Robust and Unsupervised Anomaly Discovery in Dynamic Networks, *27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

5. Wen-Ting Chung, Yu-Ru Lin, Ang Li, **Ali Mert Ertugrul**, Muheng Yan. March with and without Feet: The Talking about Protests and Beyond, *10th International Conference on Social Informatics (SocInfo)*, 2018.

6. Roberto Napoli, **Ali Mert Ertugrul**, Alessandro Bozzon, Marco Brambilla. A User Modeling Pipeline for Studying Polarized Political Events in Social Media, *KDWEB Workshop, co-located with ICWE 2018*, 2018.

7. **Ali Mert Ertugrul**, Pinar Karagoz. Movie Genre Classification from Plot Summaries using Bidirectional LSTM, *12th IEEE International Conference on Semantic Computing (ICSC)*, 2018.

8. **Ali Mert Ertugrul**, Burak Velioglu, Pinar Karagoz. Word Embedding based Event Detection on Social Media, *12th International Conference on Hybrid Artificial Intelligence Conference (HAIS)*, 2017.

9. **Ali Mert Ertugrul**, Itir Onal, Cengiz Acarturk. Does the strength of sentiment matter? A regression based approach on Turkish social media. *International Conference on Natural Language & Information Systems (NLDB)*, 2017.

10. **Ali Mert Ertugrul**, Onur Demirors. A Method for Modeling Business Processes in a Role-based and Decentralized Way, *8th International Conference on Subject-Oriented Business Process Management (S-BPM)*, 2016.

11. **Ali Mert Ertugrul**, Onur Demirors. An exploratory study on role-based collaborative business process modeling approaches, *7th International Conference on Subject-Oriented Business Process Management (S-BPM)*, 2015.

12. **Ali Mert Ertugrul**, Gokcen Yilmaz, Murat Salmanoglu, Onur Demirors. The Effect of Highlighting Error Categories in FSM Training on the Accuracy of Measurement, *Joint Conference of International Workshop on Software Measurement and the International Conference on Software Process and Product Measurement (IWSM-MENSURA)*, 2014.

13. Itir Onal, **Ali Mert Ertugrul**, Ruken Cakici. Effect of Using Regression on Class Confidence Scores in Sentiment Analysis of Twitter Data, *5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA), co-located with ACL 2014*, 2014.

14. **Ali Mert Ertugrul**, Itir Onal. RemindMe: An Enhanced Mobile Location-Based Reminder Application, *International Conference on Future Internet of Things and Cloud (FiCloud)*, 2014.

15. Itir Onal, **Ali Mert Ertugrul**. Effect of Using Regression in Sentiment Analysis, *22nd IEEE Conference on Signal Processing and Communications Applications (SIU)*, 2014.

16. Ozan Rasit Yurum, Ozden Ozcan Top, **Ali Mert Ertugrul**, Onur Demirors. Yazılım Süreç Değerlendirme Araçlarının Karşılaştırılması: Bir Çoklu Durum Çalışması, *8. Ulusal Yazılım Mühendisliği Sempozyumu (UYMS)*, 2014.

17. **Ali Mert Ertugrul**, Itir Onal. Çeşitli Konum Etiketleme Opsiyonlarıyla Zenginleştirilmiş Yeni Bir Konum Bazlı Hatırlatma Uygulaması, *8. Ulusal Yazılım Mühendisliği Sempozyumu (UYMS)*, 2014.