Modeling Information Precursors for Event Forecasting

Yue Ning

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

Doctor of Philosophy
in
Computer Science and Applications

Naren Ramakrishnan, Chair
Chang-Tien Lu
Chris North
Huzefa Rangwala
Chandan Reddy

June 7th, 2018
Arlington, Virginia

Keywords: Information Reciprocity, Precursor Learning, Event Modeling, Event Forecasting.

Copyright 2018, Yue Ning
Modeling Information Precursors for Event Forecasting

Yue Ning

(ABSTRACT)

This dissertation is focused on the design and evaluation of machine learning algorithms for modeling information precursors for use in event modeling and forecasting. Given an online stream of information (e.g., news articles, social media postings), how can we model and understand how events unfold, how they influence each other, and how they can act as determinants of future events?

First, we study information reciprocity in joint news and social media streams to capture how events evolve. We present an online story chaining algorithm which links related news articles together in a low complexity manner and a mechanism to classify the interaction between a news article and social media (Twitter) activity into four categories. This is followed by identification of major information sources for a given story chain based on the interaction states of news and Twitter. We demonstrate through this study that Twitter as a social network platform serves as a fast way to draw attention from the public to many social events such as sports, whereas news media is quicker to report events regarding political, economical, and business issues.

In the second problem we focus on forecasting and understanding large-scale societal events from open source datasets. Our goal here is to develop algorithms that can automatically reconstruct precursors to societal events. We develop a nested framework involving multi-instance learning for mining precursors by harnessing temporal constraints. We evaluate the proposed model for various event categories in multiple geo-locations with comprehensive experiments.

Next, to reinforce the fact that events are typically inter-connected and influenced by events in other locations, we develop an approach that creates personalized models for exploring spatio-temporal event correlations; this approach also helps tackle data/label sparsity problems across geolocations.

Finally, this dissertation demonstrates how our algorithms can be used to study key characteristics of mass events such as protests. Some mass gatherings run the risk of turning violent, causing damage to both property and people. We propose a tailored solution for uncovering triggers from both news media and social media for violent event analysis.
This work was partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337, the Office of Naval Research under contract N00014-16-C-1054, and the U.S. Department of Homeland Security under Grant Award Number 2017-ST-061-CINA01. The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF, IARPA, DoI/NBC, or the US Government.
Modeling Information Precursors for Event Forecasting

Yue Ning

(GENERAL AUDIENCE ABSTRACT)

Today, massive open source information is widely available through news and social media, but analyzing this information is a complex task. It is imperative to develop algorithms that can automatically reconstruct the clues to societal events that are reported in news or social media. The focus of this dissertation is on simultaneously uncovering precursors to societal events and using such precursors to forecast upcoming events. We develop various machine learning algorithms that can model event-related data and determine the key happenings prior to an event that have the greatest predictability to such events in the future. We use our algorithms to understand the nature of precursors to civil unrest events (protests, strikes, and ‘occupy’ events) and why some of these events turn violent.
Acknowledgments

I am indebted to the assistance of my thesis committee: Dr. Naren Ramakrishnan, Dr. Chang-Tien Lu, Dr. Chris North, Dr. Huzefa Rangwala, and Dr. Chandan K. Reddy. Each of them have remained supportive of my efforts and the desire to see this through.

My deepest gratitude it to my Ph.D. research advisor, Dr. Naren Ramakrishnan, for his generous financial support and invaluable academic guidance throughout my Ph.D. work. To say I learned a great deal from Naren Ramakrishnan would be an understatement. He not only kept me on the path and believed in me when I didn’t, but also encouraged me and provided much valuable advice on my career. I will never forget the time and effort that he has invested to make my graduate study productive and successful.

Dr. Chang-Tien Lu is a person who has been very caring and supportive with his valuable suggestions when I have approached him in times of need. He has also provided helpful advice on teaching for my future career. I am obliged to Dr. Chris North who has been very patient and provided insightful comments and suggestions for my preliminary proposal through my research defense and my final defense. Dr. Huzefa Rangwala, who initiated me into a few research projects, deserves special mention as he has been an important mentor all along the line. He has spent hours with me discussing technical problems and advising me on writings and presentations. His concern and care will always have me in his debt. My thanks to Dr. Chandan K. Reddy for his valuable advice into many aspects of my research and career path. He shared with me his feedback and skills on topics from research problem formulation to giving a professional talk in public.
Without their guidance, it would have been much more difficult, or even impossible, to accomplish as much as I have.

I want to thank Dr. Ravi Tandon who helped me a lot during my early graduate studies at Virginia Tech. His shared experiences and suggestions on producing research works and making impressive presentations are a true treasure to me.

My thanks to James Starz who has collaborated with me for the past few years and provided invaluable knowledge and data to my research projects.

The members and alumni of Discovery Analytics Center have contributed to my research work and academic life at Virginia Tech immensely. I am truly grateful for my colleagues: Sathappan Muthiah, Rongrong Tao, Rupinder Paul, Dr. Hao Wu, Dr. Wei Wang, Dr. Huijuan Shao, Dr. Fang Jin, Dr. Prithwish Chakraborty, Dr. K.S.M Tozammel Hossain, Dr. Saurav Ghosh, Ping Wang, Dr. Tian Shi, Dr. Ting Hua, Dr. Liang Zhao, Dr. Parang Saraf, Yaser Keneshloo, Juanita Victoria, Wanawsha Hawrami, and Joyce Newberry.

I also want to show my great appreciation of my fiancé, Jeffrey, whose support and love allowed me to completely devote myself to my Ph.D. research.

At last, I want to thank my parents. They raised me, understood me, and encouraged me in the pursuit of my Ph.D. degree.

Y. N.
Arlington
To my parents.
# Contents

1 Introduction .................................................. 1

1.1 Objectives and Research Questions .............................. 1
   1.1.1 Information Reciprocity between News Media and Social Media .......... 2
   1.1.2 Precursor Mining for Event Forecasting .............................. 2
   1.1.3 Spatio-Temporal Event Correlation .................................. 3
   1.1.4 Uncovering Triggers for Violent Events from Media Sources ............ 3

1.2 Organization of the Dissertation ................................ 3

2 Background .................................................. 5

2.1 Information Reciprocity ....................................... 5

2.2 Event Forecasting and Precursor Mining .......................... 7

2.3 Spatio-Temporal Precursor Learning .............................. 9

3 Information Reciprocity between News and Social Media ......... 10

3.1 Background .................................................. 11

3.2 Methodology .................................................. 13
3.2.1 Story Chaining of News Articles ........................................... 13
3.2.2 Retrieval of tweets related to news ........................................... 16
3.2.3 Identifying Interaction Patterns between News and Tweets ............ 17
3.2.4 Clustering of Interaction Patterns ............................................ 20
3.2.5 Topic Modeling ................................................................. 21
3.3 Dataset and Experiments ......................................................... 22
3.4 Results ................................................................................. 23
3.5 Summary ................................................................................. 29

4 Multi-Instance Learning for Mining Precursors .................................. 30
4.1 Background ............................................................................ 31
4.2 Problem Formulation ............................................................... 33
  4.2.1 Formal Definition and Notation ............................................... 35
4.3 Methodology .......................................................................... 36
  4.3.1 Nested MIL model (nMIL) ..................................................... 37
  4.3.2 Sequential Model (nMILΩ) ..................................................... 39
  4.3.3 Multiclass Classification ...................................................... 39
  4.3.4 Optimization ..................................................................... 40
  4.3.5 Precursor discovery using nMIL ............................................ 41
4.4 Experiments ............................................................................ 42
  4.4.1 Datasets ............................................................................. 42
  4.4.2 Experimental Protocol ......................................................... 43
  4.4.3 Comparative Approaches .................................................... 44
List of Figures

2.1  Comparison of multi-instance learning (left) and supervised learning (right) . . . . . .  8

3.1  An example of a story chain of news articles from Brazil. . . . . . . . . . . . . . . .  12

3.2  A schematic framework of the proposed interaction pattern mining. . . . . . . . . . . .  13

3.3  Twitter peaks corresponding to one news article . . . . . . . . . . . . . . . . . . . . .  18

3.4  The geometric interpretation of four interaction types. . . . . . . . . . . . . . . . . . .  20

3.5  Population and event type distribution in GSR chains . . . . . . . . . . . . . . . . . .  23

3.6  Difference in event type and population distributions for different interaction pattern cluster. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .  25

3.7  Topic distributions of each interaction pattern cluster. . . . . . . . . . . . . . . . . . .  27

4.1  An example of precursor story line for a protest event in Argentina. . . . . . . . . .  31

4.2  An overview of the proposed nMIL model. . . . . . . . . . . . . . . . . . . . . . . . . . .  36

4.3  F1 score evaluation on three countries. . . . . . . . . . . . . . . . . . . . . . . . . . .  46

4.4  Estimated probabilities for negative examples and positive examples in the nMIL model.  49

4.5  Mean of relative cosine values w.r.t target events in history days in the nMIL model.  49

4.6  The distribution of relative cosine similarity and relative entity hit score. . . . . . .  50
4.7 Case studies of precursor storyline in Argentina and Mexico discovered by the nMIL model ........................................... 51
4.8 Sensitivity analysis on $\beta$ and $\lambda$ in the nMIL model ........................................ 53
5.1 Precursor event sequence discovered by our method for a protest event .............. 56
5.2 The proposed system framework of STAPLE ....................................................... 58
5.3 An example of the spatio-temporal correlation graph ......................................... 61
5.4 Learning representations. Top portion corresponds to the document embeddings and the bottom portion corresponds to the entity embeddings .......................... 62
5.5 Prediction performance (Recall, Precision, and F1 scores) on six datasets for the comparison methods ...................................................... 69
5.6 F1 lift per city using STAPLE compared to state-of-the-art model, Nested ........... 70
5.7 Comparison of the average normalized Jaccard index for precursor news and non-precursor news articles ......................................................... 72
5.8 Test accuracy for two countries over leadtime ranging from 1 to 5 days ............ 73
5.9 Sensitivity analysis on hyper-parameters $\lambda_1$ and $\lambda_2$ ($= \lambda_3$) ................. 74
6.1 Average number of gatherings (in prior week) that precede violent and non-violent events ................................................................. 76
6.2 Day module of the proposed method ................................................................. 78
6.3 The proposed method of CrowdForecaster ....................................................... 78
6.4 AUC scores of models for the Argentina dataset ................................................ 88
6.5 Computation time of models for the Argentina dataset ..................................... 88
6.6 Sensitivity analysis of model parameters ......................................................... 90
List of Tables

3.1 Statistical properties of GSR and Non-GSR chains. .......................... 23
3.2 Percentage of Twitter, News starts for GSR story-chains. .................... 23
3.3 Important words in topics inferred by LDA. ..................................... 26
3.4 Top topics for clusters. ............................................................... 26
3.5 Story Chains with Interaction Patterns and Main Influencer. ............... 28
4.1 Notations used in the nMIL model .................................................. 34
4.2 Event forecasting performance comparison. ...................................... 45
4.3 F1-measure for rMIL^{avg} and nMIL^{Δ} models with varying leadtime. .... 48
4.4 Multi-Class F1-Measure for rMIL^{avg} and nMIL models. .................... 52
5.1 Datasets used in our experiments .................................................... 66
5.2 F1 evaluation for the proposed methods ........................................... 68
5.3 Automatically discovered precursor news stories and relevant entities for one protest event of interest ......................................................... 71
6.1 Violent event forecasting performance comparison based on Accuracy (Acc) and AUC score w.r.t to state-of-the-art methods ......................... 84
6.2 AUC scores of the proposed method and the best baseline with lead time from 1 to 4 for violent events ................................................................. 85
6.3 Quality scores for the proposed method and the delivery from online system, EMBERS [46] ................................................................. 86
6.4 Precursor events and word distributions in past seven days for violent protest events and one non-violent protest event ......................... 92
6.5 Top Twitter words in countries: Argentina, Brazil, and Columbia ............. 93
6.6 Top Twitter words in countries: Paraguay and Venezuela ......................... 94
Chapter 1

Introduction

1.1 Objectives and Research Questions

With the rapid growth in online open source information such as news, blogs, and social media, a key topic of research interest is to utilize this vast assortment of data to support decision making processes and forecasting of upcoming events. In past research, open source data (e.g., social media and news feeds) have been proven to serve as successful surrogates in forecasting a broad class of events such as disease outbreaks, elections, and stock markets. In this work, we focus on the domain of forecasting societal uprisings such as civil unrest movements.

While many past studies focus on predictive accuracy, in practice many scenarios require the understanding of evidential causalities for events of interest. Given the immense potential need for understanding event progression, it is important to model how information is reported in news and social media with a view to understanding event evolution and precursors for events of interest.

In this work, four research questions related to event forecasting and understanding are studied. We focus on modeling societal events and their precursors with multiple constraints: (1) discovering information interaction pattern between news and social media along event storylines. (2) developing machine learning formulations for mining event precursors and forecasting events. (3) modeling
spatio-temporal properties of event progression. (4) capturing the dynamics of when protest events turn violent.

1.1.1 Information Reciprocity between News Media and Social Media

In recent years, the amount of information shared (both implicit and explicit) between traditional news media and social media sources like Twitter has grown at a prolific rate. Traditional news media is dependent on social media to help identify emerging developments; social media is dependent on news media to supply information in certain categories. However, very little work has been done to comprehensively analyze the information interaction patterns between them. It is anticipated that the in-depth analysis of understanding their symbiotic relationship could help to find how topics flow among different media for stories. How can we systematize the interaction patterns in story chains? How do we decide an ideal utilization of information sources for the event developments in a specific domain or topic? Research in information reciprocity of media provides help in maximizing the usage of different media according to the tasks. For instance, traditional news and social media share a symbiotic relationship. For some topics such as political events, news media tends to report developments rapidly; while in some other cases, it is dependent on social media to help spread its news to the masses.

1.1.2 Precursor Mining for Event Forecasting

Modeling precursors for event forecasting is a costly challenge because it is extremely time-consuming to manually label the precursors. How can we learn a system to complete the two tasks at the same time (forecasting events in future and identifying precursor for these events) in a semi-supervised environment? A naive way would be providing documents with higher semantic similarities to the target events. However, the text description is not always offered for the events and precursors do not necessarily have higher semantic similarity to the target events. A natural fit would be selecting documents that are recognized as higher predictive indicators for the target events using a framework such as multi-instance learning.
1.1.3 Spatio-Temporal Event Correlation

A key challenge that remains in event forecasting and understanding is the incorporation of spatial and temporal correlations inherent in large-scale societal event occurrences. The absence of such modeling creates shortcomings in not just quality of inference, but also curtails interpretation by human analysts. Observing event reports for different geolocations, there is a strong spatio-temporal correlation for event occurrences. With a lack of labels for precursors and data sparsity for certain geolocations, how can we design a framework for multiple geolocations to support insufficient data, imbalanced class distributions, and partial labels? Personalized models are proven to achieve better performances compared to a global model when facing these challenges. A natural framework here is multi-task learning where different learning tasks share common features and are influenced by each other.

1.1.4 Uncovering Triggers for Violent Events from Media Sources

Mass gatherings involved in civil disobedience activities run the risk of turning violent, causing damage to both property and people. While civil unrest is a rather common phenomenon, only a small subset of them involve crowds turning violent. With sparse datasets, how can we better capture signals and distinguish which events are likely to lead to violence?

1.2 Organization of the Dissertation

The dissertation is organized as follows.

In Chapter 2, we discuss related research pertaining to the problems of information reciprocity, event forecasting, precursor mining, and spatio-temporal precursors.

In Chapter 3, we undertake our study of information reciprocity underlying story chains in two heterogeneous text media: traditional news media and social media sources like Twitter. Our study shows that Twitter as a social network platform serves as a fast way to draw attention from the public to many social events such as sports news whereas news media is quicker to report
events regarding political, economical, and business issues. Certain topics were found to have similar influence from both Twitter and news media.

In Chapter 4, we study the problem of precursor learning and event forecasting. Specifically, given a collection of streaming news articles from multiple sources we develop a nested multiple instance learning approach to forecast significant societal events such as protests. Using data from three countries in Latin America, we demonstrate how our approach is able to consistently identify news articles considered as precursors for protests. Our empirical evaluation demonstrates the strengths of our proposed approach in filtering candidate precursors, in forecasting the occurrence of events with a lead time advantage, and in accurately predicting the characteristics of civil unrest events.

In Chapter 5, we investigate the spatio-temporal event correlations. Given a geolocation (city) and its event reports for a time period, we present a novel multi-task spatio-temporal correlation graph model within a multi-instance learning framework that tackles the problem of scarce data distribution and reinforces multiple geolocation precursors with augmented representations. Through studies of civil unrest movements in numerous countries, we demonstrate the effectiveness of the proposed model for precursor discovery and event forecasting.

In Chapter 6, we design a tailored multi-instance learning structure for forecasting a sparse event category, viz. violent protests. Using data from five countries in Latin America, we demonstrate not just the predictive utility of our approach, but also its effectiveness in discovering triggering factors, especially in uncovering how and when crowd behavior begets violence.

In Chapter 7, we summarize our work and discuss future directions.
Chapter 2

Background

Background work relevant to this dissertation can be categorized into a few groups: (i) *Storytelling and information interactions*. Storytelling aims to relate given start and end documents by uncovering a series of intermediate documents. Information interaction patterns focus on the topical analysis over time of interactions between different media sources. (ii) *Event detection and forecasting*. Event detection and forecasting from online open source datasets have been an active area of research in the past decade. Both supervised and unsupervised machine learning techniques have been developed to tackle different challenges. (iii) *Precursor learning*. Identifying precursors for significant events is an interesting topic and has been used extensively for interpretive narrative generation and in storytelling algorithms. (iv) *Spatio-temporal event modeling*. Multi-task learning and transfer learning have been studied to explore connections in multiple geolocations involving temporal constraints and multiple spatial resolutions.

2.1 Information Reciprocity

*Storytelling* (or “connecting the dots”) was introduced by Kumar et.al. in [30] as a concept in knowledge discovery and data analytics. It aims to relate given start and end documents by uncovering a series of intermediate documents. This problem has been studied in entity networks.
social networks, cellular networks, document collections. Most existing approaches to storytelling use offline data wherein a user must specify the start and end documents of the chain and the algorithm aims to uncover the sequence of relationships between the two endpoints. Shahaf et al. in define concepts of chain coherence, coverage, and connectivity offering insights into the storytelling process. This approach relies on building bipartite word-document or word cluster graphs making it computationally expensive. Leskovec et al. develop a meme-tracking approach for online text and observe a “heartbeat”-like pattern in the handoff between news and blogs.

Twitter’s role in event reporting and as a news source is well established. Sakaki et al. used Twitter users as sensors to estimate locations of events such as earthquakes. Chierichetti et al. analyzed tweet streams to identify important events and the tweet production/consumption patterns around the key events. They observed a robust “heartbeat” phenomenon when key events happen. Ramakrishnan et al. use Twitter with other data sources to forecast protests and other civil unrest events. Y. Hu et al. present a joint Bayesian model framework called ET-LDA to extract topics covered by the event and to perform event segmentation in one unified framework. Jin et al. proposed a topic model which learns topic distributions for two datasets by transferring topical knowledge.

Communication patterns of a social network within itself and with external platforms have been explored with diverse techniques. Hopcroft et al. have studied the reciprocal relationship in a dynamic social network and their findings suggest how individuals’ behavior are determined by social structures. There have been studies of the relationship between Twitter and traditional news media and especially how fast and powerful Twitter can be for publishing or discussing live stories. Also, the role of Twitter in news reporting has been explored. Petrovic et al. examine the extent to which news reports and Twitter overlap and whether Twitter often reports faster by manually identifying major news events. Kwak et al. studied the topological characteristics of Twitter as a platform for information sharing. Regarding connecting tweets to news, Sankaranarayanan et al. developed a news processing system called TwitterStand to capture tweets that correspond to late breaking news.
The above efforts chip away at the problem of modeling the interaction between news and social media but only address partially our goals here. They either focus on how news articles can be chained together to study news-news interaction or study how Twitter can replace news. At the other extreme, while studies such as Petrovic et al. [45] look at the overlap between news and Twitter, these works require significant human involvement. In our framework, we combine temporal dynamic characteristics of tweets and align them to news articles for each story. We define interaction patterns in both quantitative and qualitative ways for story chains and cluster chains to infer topical similarities in Chapter 3.

### 2.2 Event Forecasting and Precursor Mining

**Event Detection and Forecasting**  Event detection and forecasting from online open source datasets has been an active area of research in the past decade. Both supervised and unsupervised machine learning techniques have been developed to tackle different challenges. To estimate the time for future events, linear regression models have been proposed using simple features [4, 7, 19, 44]. Advanced techniques use a combination of sophisticated features such as topic related keywords, as input to support vector machines, LASSO and multi-task learning approaches [64, 49]. Ramakrishnan et al. [46] designed a framework (EMBERS) for predicting civil unrest events in different locations by using a wide combination of models with heterogeneous input sources ranging from social media to satellite images. Zhao et al. [71] combine multi-task learning and dynamic features from social networks for spatial-temporal event forecasting. Generative models have also been used in [70] to jointly model the temporal evolution in semantics and geographical burstiness within social media content. Laxman et al. [32] designed a generative model for categorical event prediction in event streams using frequent episodes. However, few existing approaches provide evidence and interpretive analysis as support for event forecasting.

**Identifying Precursors**  Identifying precursors for significant events is an interesting topic and has been used extensively for interpretive narrative generation and in storytelling algorithms [22]. Rong et al. [50] developed a combinational mixed Poisson process (CMPP) model to learn social, external and intrinsic influence in social networks.
Multiple Instance Learning  In the multiple instance learning (MIL) paradigm as shown in Figure 2.1, we are given labels for sets of instances commonly referred as bags or groups. However, individual instance-level labels are unknown or missing. The bag-level labels are assumed to be an association function (e.g., OR, average) of the unknown instance level labels. One approach to MIL adapts support vector machines (SVMs) by: (i) modifying the maximum margin formulation to discriminate between bags rather than individual instances [3], and (ii) developing kernel functions that operate directly on bags [16]. Other multiple instance learning approaches and various applications are found in a detailed survey [2]. Specifically, the generalized MIL [67] formulation assumes the presence of multiple concepts and a bag is classified as positive if there exists instances from every concept. Relevant to our work, besides predicting bag labels, Liu et al. [37] seek to identify the key instances within the positively-labeled bags using nearest neighbor techniques. Recent work [29] has focused on instance-level predictions from group labels (GICF) and allowed for the application of general aggregation functions with applications to detecting sentiments associated with sentences within reviews.

We propose a framework in Chapter 4, which can be viewed as complementary to prior work, casting the forecasting and precursor discovery problems via novel extensions of multiple instance learning.
2.3 Spatio-Temporal Precursor Learning

**Multi-task learning**  Multi-task learning (MTL) [9, 14] considers multiple related tasks together to improve generalization performance by leveraging the domain-specific information inherent in the training signals of related tasks. It has been applied successfully across many application domains such as computer vision [17], bioinformatics [47], and natural language processing [11]. Many multi-task learning approaches have been proposed for spatio-temporal event detection [71]. To the best of our knowledge, our work is the first time MTL is used to develop a framework for civil unrest event precursor discovery. In our proposed method, we study a multi-task learning framework considering temporal constraints among events to detect spatio-temporal precursors while predicting events of interest.

**Representation Learning**  In practice, finding good feature representations to model news articles is not a trivial problem. Traditional bag-of-words representations allow for easy interpretation but require preprocessing and feature selection. Several researchers have developed efficient and effective neural network based language models [5, 40]. Entity recognition has also been widely applied in natural language processing tasks [36, 15]. Most societal events are related to or even caused by known entities such as persons, organizations, and geolocations.

In Chapter 5, we cast the spatial event forecasting and precursor discovery problems in a multi-task learning framework with a fusion penalty based on spatio-temporal correlation graphs and augmented representations. We combine document level embeddings with recognized entity embeddings for analyze the impact of significant names or locations on the event development.
Chapter 3

Information Reciprocity between News and Social Media

Analyzing the information reciprocity between different media for event development helps researchers better decide the usage of multiple source datasets. In recent years, the amount of information shared (both implicit and explicit) between traditional news media and social media sources like Twitter has grown at a prolific rate. Traditional news media is dependent on social media to help identify emerging developments; social media is dependent on news media to supply information in certain categories. In this chapter, we present a principled framework for understanding their symbiotic relationship, with the goal of (1) understanding the type of information flow between news articles and the Twitterverse by classifying it into four states; (2) chaining similar news articles together to form story chains and extracting interaction patterns for each story chain in terms of interaction states of news articles in the story chain, and (3) identifying major interaction patterns by clustering story chains and understanding their differences by identifying main topics of interest within such clusters.
3.1 Background

Social media sources like Twitter, Facebook, Instagram, Reddit etc. have grown to become an effective part of one’s daily life. Twitter has emerged as a powerful medium where people report and comment on everyday happenings. With the proliferation of social media, information shared in traditional media sources like news and blogs is no longer independent of the information in social media; there is implicit information exchange across them. Twitter for instance, tends to break developments rapidly for events that involve mass public involvement such as sporting events and natural disasters [52]; news on the other hand is still the prime source for events related to politics and government.

Traditional and social media sources thus share a symbiotic relationship. In many scenarios, traditional news media is dependent on social media to help spread its news to the masses whereas in other scenarios, social media is dependent on traditional media to supply new information to comment/feed upon. Such interdependencies tend to vary based on the popularity of a topic in social media and also on the geographic location of the topic. In this chapter we try to uncover such symbiotic relationships through a principled framework by identification of interaction patterns between news and tweets, understanding the differences in such interaction patterns, and imputing such differences corresponding to distinct information topics.

To illustrate an example interaction pattern, Fig. 3.1 showcases a series of news reports following a fire accident at a nightclub in Santa Maria, Brazil on 27th Jan 2013 as detected by our framework. This figure also depicts trends in Twitter with respect to keywords and actors (persons and organizations) mentioned in news reports. From the Twitter trends, we can see that for certain news reports there is a peak in the corresponding Twitter activity profile before its publication time, whereas for some, such peaks happen after the news report’s publication. This observation suggests that we can use such timing and volume information to capture the direction of information flow between news and Twitter. For instance, in Fig. 3.1, the newswire breaks the story first. This news was possibly captured by Twitter next as there is a spike in Twitter activity before the second news article. News then immediately follows up and this way both news and Twitter reciprocate. Throughout this story chain progression, interactions between news and Twitter activity are clearly evident and mining this interplay and reciprocity is the goal of the framework proposed.
Figure 3.1: An example of a story chain from Brazil about a fire accident at a nightclub. For every news report (circles), its corresponding twitter activity profile is shown. The activity profiles are centered around the article publication time and direction of arrow from news reports to the activity profile indicate the direction of information flow and thereby interaction type. Observe the multiplicity of directionality in this example.

in this chapter. To the best of our knowledge, this is one of the first approaches to do such a study. We next summarize the main contributions:

- We present an online story chaining algorithm which *chains* related news articles together in a low complexity manner. Our algorithm is based on weighted scores of similarities across news articles for three sets of features: textual features (related to keywords), spatial features (such as locations and geographical coordinates), and actors (such as person(s), and organizations mentioned in the articles).

- We introduce a mechanism to classify the interaction between a news article and Twitter activity around its publication time through four *interaction states*: $N$ (information flow from news article to Twitter), $T$ (information flow from Twitter to news article), $B$ (bi-directional interaction between news and Twitter), and $E$ (empty, or no interaction). This encoding mechanism is applied to all articles in a story chain resulting in a string of interaction states; and the collective string is the *interaction pattern* of a story chain.

- We identify the major source of information for a given story chain based on the interaction states of every news report in the chain and its corresponding quantitative weights. To this end, the interaction patterns of story chains are used to identify distinctive clusters
of interactions. Distinct clusters of interaction patterns are further studied to check for clear and explicit dissimilarities in terms of the content reported by the news articles in each cluster. LDA-based topic modeling is used to discern content differences between the different interaction pattern clusters.

### 3.2 Methodology

Our overall framework (Figure 3.2) has the following main components: we first thread news articles into story chains, retrieve Twitter trends for every news report in a given story chain, detect and encode interaction patterns and finally use clustering and topic modeling to understand the topical differences among different interaction patterns. Each one of these components is described in detail in the following sections.

#### 3.2.1 Story Chaining of News Articles

The chaining methodology is developed with the goal of identifying all documents related to a news story and to keep track of the news story as new documents arrive. Documents belonging to such a chain cover the same event and are ordered by time. Traditional clustering approaches can cluster together documents about similar events but are insufficient to separate out documents of each
individual event. Thus we formalize an approach that chains together documents about an event as they appear, in order to build a narrative thread of that event. The algorithm operates in an incremental fashion wherein every new input article is analyzed as it arrives and is appended to already existing chain(s). This analysis involves a two-step process. In the first step, we compare an incoming article $D_i$ to articles from the last $n^1$ days to identify the most similar articles and then designate candidate chains to which the current article can be attached to. If no similar articles are found, then a new chain is created with this article as a seed. Further, to assess if two documents are referring to the same underlying context, we calculate their *similarity scores* with respect to three features:

- textual features, denoted by $T(D_i)$,
- spatial features, denoted by $L(D_i)$, and
- actors, denoted by $A(D_i)$.

The textual features are represented by the TF-IDF vector of the tokens present in the document. The spatial features are the set of locations mentioned within the text of an article. Every phrase/token identified as a location name by a named entity recognizer (NER) is resolved into a $<$country, state, city$>$ tuple with help of a geocoder based on probabilistic soft logic [28]. Details of the geocoding methodology are described in our prior work [42]. The weight $l_{ij}$ of a location entity in the spatial feature vector represents the probability of appearance of location $l_j$ in the document $D_i$.

$$l_{ij} = \frac{\text{Frequency of } l_j \text{ in } D_i}{\sum_k \text{Frequency of } l_k \text{ in } D_i}.$$  

(3.1)

Similarly, the actors feature vector ($A(D_i)$) represents the set of actors (persons, organizations as detected by NER) mentioned in an article. The weight $a_{ij}$ of each element in the actors feature vector represents the probability of appearance of actor $a_j$ in document $D_i$ and is defined similar to (3.1).

---

1Empirically, $n = 14$ (2 weeks) was found to be most effective.

2http://www.basistech.com/text-analytics/rosette/
The total weighted similarity measure between two documents, $D_i$ and $D_j$, is then defined as follows

$$\text{sim}(D_i, D_j) \triangleq \alpha f(T(D_i), T(D_j)) + \beta f(\mathcal{L}(D_i), \mathcal{L}(D_j)) + \eta f(\mathcal{A}(D_i), \mathcal{A}(D_j)),$$

(3.2)

where $f$ denotes a similarity metric such as cosine similarity or Jaccard’s coefficient and the weighting coefficients $\alpha, \beta, \eta$ are chosen such that $\alpha + \beta + \eta = 1$. The textual similarity in this equation captures the similarity in terms of topical content whereas the spatial and actor vectors capture the similarity in terms of the event(s) described in the two documents. Thus the choice of the weights $\alpha, \beta, \eta$ control the relative coherence of two documents w.r.t. textual, spatial, and actor related features.

Eqn. 3.2 is used to obtain articles most similar to the current article from the past $n$ days and thus a set of candidate chains to which the current article can attach could also be found. Once a candidate set of chains are found, in the second step, the candidate set is pruned based on the coherence of the article $D_i$ with a story chain $C_j$. Here, coherence is calculated as the weighted sum of coherence between the spatial and actor feature vectors of an article and the spatial and actor feature vectors of a chain. The spatial feature vector $\mathcal{L}(C_j)$ and the actor feature vector $\mathcal{A}(C_j)$ of a chain are defined similar to (3.1) by considering all news articles in the chain as a single document.

The coherence between a chain $C_j$ and document $D_i$ is defined as

$$\text{coh}(D_i, C_j) = \theta g(\mathcal{L}(D_i), \mathcal{L}(C_j)) + \phi g(\mathcal{A}(D_i), \mathcal{A}(C_j)),$$

where $g$ is any similarity measure and the coefficients $\theta, \phi$ are chosen such that $\theta + \phi = 1$. The spatial and actor feature vectors for a chain are then updated every iteration if there is any update i.e., any new document is added to the end of the chain.

The article $D_i$ is attached at the end of all chains such that $\text{coh}(D_i, C_j) \geq \Gamma$, where $\Gamma$ is the threshold which is used to tradeoff chain length and coherence. A higher value of $\Gamma$ will cause the chains to be shorter but more coherent, and vice versa. If no chain passes similarity threshold $\Gamma$, then a new chain is created with this article. This two step process is repeated for every new article. Jason et al. present an evaluation of this chaining methodology in [55].
3.2.2 Retrieval of tweets related to news

Retrieving tweets related to a given news article is not trivial as tweets are comparatively very short (only 140 chars, now 280 chars) and it is also necessary to find tweets associated with a news article both before and after its publication time in order to understand the information flow between news and Twitter. Sometimes, tweets mention the shortened URL of the actual news article thereby establishing an explicit connection indicating flow of information from traditional news media to the Twitterverse. However, such tweets are very few in number. On an average, from our experiments, we found that for a given news article, in our collection only about 5-6 tweets explicitly mention its URL (both shortened and unshortened forms were considered). On the other hand, certain news articles do also cite tweets or Twitter usernames and hashtags in their content which can be used to find associated tweets appearing before the article got published. This again is only a handful. Therefore, given the limitations of the API we used (Topsy), we resort to techniques of obtaining twitter count metrics by identifying tweets by keywords instead of techniques like topic filtering, BM25 and Rocchio methods. Specifically we follow a four-step process as illustrated below:

Step 1: Collect tweets mentioning a given URL. We harvest both the mentions of shortened and unshortened forms of the given URL.

Step 2: Extract top 10 keywords from the list of keywords obtained after tokenization and stopword removal of the text of tweets obtained in the earlier step. This list of keywords is combined with a set of entity words obtained by performing language enrichment on the news article (see Section 3.2.1).

Step 3: Remove items from the previously obtained list of keywords plus those entities that are common to other articles in the same chain as the current article. This is necessary because all news articles in a story-chain share some common topics and so will their corresponding Twitter activity. Thus it is important that we extract information unique to a particular news report versus that of other news reports in a chain. This step is necessary as it helps us study news-Twitter interaction at an individual article levels without having to consider inter-dependencies.

Step 4: Download hourly count metrics for each element in the keywords plus in the entities list obtained in the last step. The Twitter count metric download is limited to the time window of \([t_0 - 7, t_0 + 7]\) days, where \(t_0\) is the article publication date.
### 3.2.3 Identifying Interaction Patterns between News and Tweets

At this point, we have news articles grouped together to form story chains and for each news report in a story chain, we have its corresponding Twitter activity profile. In this section, we discuss how interaction is defined for a single news article and then use this information to define interaction patterns for a whole chain.

Peaks in Twitter activity showcase interestingness and can indicate either inflow of information from another source or possible triggers for outflow of information to a different source. We assume that the presence of peaks in the Twitter activity is a good milestone to use to posit interactions between news and Twitter. For all our experiments, we assume the interaction is only between news wires and tweets and that there is no other third source. The algorithm for peak detection [13] is detailed in Algorithm 1. Peaks are defined to be those points in time where the corresponding value is higher than its immediate surrounding (±3 hours) and the difference is much higher than the standard deviation of the entire series.

**Algorithm 1** Peak Detection in Twitter activity profile

1. procedure **DetectPeaks**(y, threshold)
2. \[ m = \text{std}(y); \ y_m = \text{mean}(y) \]
3. IND = [] as peak position array
4. for \( y_i \) in \( y \) do
5. \hspace{1cm} if \( y_i \) satisfies the following constraints then
6. \hspace{2cm} (1) \( y_i > y_{i+1} \) and \( y_i > y_{i-1} \)
7. \hspace{2cm} (2) \( y_i > y_m \)
8. \hspace{2cm} (3) \( \min(y_i - y_{i+1}, y_i - y_{i-1}) > m \)
9. \hspace{2cm} (4) \( i > \text{inds.last()} + \text{threshold} \)
10. \hspace{1cm} inds.append(i)
11. return inds

Peaks that appear close to the article publish time have higher possibility to influence or get influenced by the news article depending on whether they happens before or after the article is published. Hence, the net influence is not only based on time lag between the news article and peak but also the actual peak value. In short, we define the influence weight of a Twitter peak to be
Figure 3.3: Twitter peaks corresponding to one news article

directly proportional to its peak value and inversely proportional to the time lag between the peak and the publication time of the news article. As in Figure 3.3, the influence weights of pre- and post- article publish time peaks are summed up separately to capture the net incoming influence $W_{\text{pre}}$ and the net outgoing influence $W_{\text{post}}$ of a news article as:

$$W_{\text{pre}} = \sum_{s \in S_{\text{pre}}} \frac{v_s}{t_A - t_s}, \quad W_{\text{post}} = \sum_{s \in S_{\text{post}}} \frac{v_s}{t_s - t_A},$$ (3.3)

where

- $t_A$ is the time of publication of news article $A$.
- $S_{\text{pre}}$ is the set of peaks detected before $t_A$.
- $S_{\text{post}}$ is the set of peaks detected after $t_A$.
- $t_s$ is the occurrence time of the peak $s$ and $v_s$ is the peak value after normalizing the Twitter activity profile so that values range from 0 to 1.

Next, using the net incoming and outgoing influence weights, we define four interaction states in which a news article can be in:
N : Here, the direction of information flow is predominantly from News to Twitter. Thus, $W^{\text{pre}}$ is not significant whereas $W^{\text{post}}$ can be significantly higher compared to $W^{\text{pre}}$.

T : This state indicates Twitter is the major information source and the flow from Twitter to news is significant as compared to the reverse flow. Mathematically, $W^{\text{pre}}$ is significant and $W^{\text{post}}$ is not significantly higher than $W^{\text{pre}}$.

B : State B represents bi-directional information flow between news and Twitter. Here, both $W^{\text{pre}}$ and $W^{\text{post}}$ can be significant.

E : This state denotes absence of any significant information flow, i.e., both $W^{\text{pre}}$ and $W^{\text{post}}$ are insignificant.

Formally, these set of states are defined in Equation. 3.4.

$$\text{State}(D_i) = \begin{cases}  
N, & \text{if } W^{\text{pre}} < \rho, W^{\text{post}} \geq (1 + \lambda)W^{\text{pre}} \\
E, & \text{if } W^{\text{pre}} < \rho, W^{\text{post}} < (1 + \lambda)W^{\text{pre}} \\
T, & \text{if } W^{\text{pre}} \geq \rho, W^{\text{post}} < (1 + \lambda)W^{\text{pre}} \\
B, & \text{if } W^{\text{pre}} \geq \rho, W^{\text{post}} \geq (1 + \lambda)W^{\text{pre}} 
\end{cases}$$  

Here, $\rho$ is the significance threshold for $W^{\text{pre}}$ signifying the level of Twitter activity before the article publish time. $\lambda$ is the significance threshold for the difference between $W^{\text{pre}}$ and $W^{\text{post}}$, which corresponds to the % increase in Twitter activity after article publish time.

Fig. 3.4 shows the sectors represented by each of these four interaction types in a cartesian plane defined by $(W^{\text{pre}}, W^{\text{post}})$. This figure also shows typical Twitter activity profile(s) around article publish time corresponding to the four interaction states.

Once every article in a chain is assigned a state based on Equation. 3.4, the interaction pattern of a chain can be defined as the concatenated string of interaction states of each individual article of the story chain in temporal order. For example, for the chain in Fig. 3.1 the interaction pattern is “NTNTBNNE”. Each individual character in the string represents the interaction state of the corresponding news article with respect to its Twitter activity profile. This type of encoding is referred to as a qualitative encoding. Also every chain can be represented as a two-dimensional real valued vector where one dimension represents the $W^{\text{pre}}$ values of each article in the chain and the
other represents $W_{\text{post}}$ of each article. This form of encoding will be referred to as a *quantitative encoding*.

### 3.2.4 Clustering of Interaction Patterns

In this section, we present two approaches for clustering the story chains using their interaction patterns. Clustering is performed using both qualitative and quantitative encoding of interaction patterns as both offer different advantages.

**Clustering via qualitative encoding** – Using the qualitative encoding, every chain is represented as a string of labels (e.g., “TTEBNEB”), each label corresponding to the interaction states of articles in the chain. We then use string edit distance metrics to calculate the difference in interaction patterns among two story chains. As story chains can differ widely in lengths, we collapse repetitive letters into one to reduce the encoding size differences among different chains so that the string edit distance metrics are more effective. Thus for the example in Fig. 3.1, the collapsed representation...
is “NTNTBNE”. We explore distance based $k$-medoids clustering with a setting of 5 clusters and the following possible string edit distances:

- Levenshtein distance [35, 63], which is the edit distance between two sequences and is defined as the minimum number of single character edits to change one sequence into the other.
- Jaro-Winkler distance [25, 68], which is also a type of edit distance that was developed in the area of record linkage. It uses the idea that differences at the start of a string are more significant than edits at the end of a string.
- Ratcliff-Obershelp pattern recognition [48] which computes the similarity between two strings as the doubled number of matching characters divided by the total number of characters in the two strings.

We select the Jaro-Winkler distance for our analysis henceforth as it has a lower intra-cluster distance for $K = 5$ clusters. **Clustering via quantitative encoding** – clustering based on the qualitative encoding suffers from certain disadvantages. The primary disadvantage is that the string edit distance metrics are quite sensitive to string length. Collapsing the string encodings as we described above reduces the effect of this problem a little but does not get rid of it in its entirety. For this reason, we encode chains as a two-dimensional vector of values of $W^{pre}$ and $W^{post}$ of each individual article in the chain. This form of encoding allows us to apply multi-dimensional dynamic time warping (DTW) to help compare the differences in interaction patterns of two chains of different lengths. DTW [62] finds the optimum alignment between two sequences of observations by warping the time dimension with certain constraints, thus allowing comparison of two sequences of different lengths.

### 3.2.5 Topic Modeling

Thus far now, we clustered story chains by employing the interaction patterns between news and tweets. To identify hidden topics underlying each cluster of story chains and explore if certain specific interaction patterns show interests in specific topics, we apply topic modeling algorithms on the news report collections. Specifically, we use latent Dirichlet allocation [6] to generate distributions
over words for each topic (and also obtain the proportions of topics distributed in a document). Then we define the weights over the mixture of topics for one cluster by:

\[ C_{j,k} = \frac{\sum_{d_{ij} \in c_j} n_{d_{ij}} \theta(d_{ij}, k)}{\sum_{d_{ij}} n_{d_{ij}}}, \]

where \( n_{d_{ij}} \) refers to the frequency of \( d_i \) in cluster \( C_j \) and \( \theta(d_{ij}, k) \) refers to the topic proportions for this document.

### 3.3 Dataset and Experiments

We show the results of our framework on real data from Brazil during the period from Nov. 2012 to Sep. 2013. This period was chosen due to unusually high social media activity and news coverage around Brazilian mass protests (also known as the “Brazilian Spring”). We collected news reports in this period from top three leading news agencies in Brazil–O Globo, Estadao, and Jornal do Brasil. The news chaining algorithm proposed in Section 3.2.1 was applied to this corpus yielding 13,529 chains, out of which chains with a minimum length of 3 were considered resulting in 9933 chains. For every news report in these story chains, geo-targeted Twitter activity profiles (limited to Brazil) were collected using the Topsy api\(^3\) as described in Section 3.2.2.

In addition to this data, we also have access to a human curated list of civil unrest events that happened during this period. This list, called the Gold Standard Report (GSR) described in Ramakrishnan et al. [46], contains news reports for each event from the three major news sources. The GSR also contains human-annotated information about the type of event and the group of people protesting as shown in Fig. 3.5. We separate the set of story chains into two categories: GSR chains (containing at least one GSR reported civil unrest event) and Non-GSR chains (with no GSR reported event). As GSR chains have more information, this segregation further highlights the differences in interaction patterns in terms of event type and population, which is not available otherwise for Non-GSR chains.

\(^3\)http://api.topsy.com/doc/
3.4 Results

We present four main results as follows:

Table 3.2: % of Twitter, News starts for GSR story-chains.

<table>
<thead>
<tr>
<th>Category</th>
<th>% News starts</th>
<th>% of Twitter starts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Housing related protests</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Other (religious &amp; cultural)</td>
<td>60%</td>
<td>40%</td>
</tr>
<tr>
<td>Govt. Policies</td>
<td>23%</td>
<td>77%</td>
</tr>
<tr>
<td>Medical</td>
<td>74%</td>
<td>26%</td>
</tr>
<tr>
<td>Agriculture</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>General Population</td>
<td>30%</td>
<td>70%</td>
</tr>
</tbody>
</table>

Figure 3.5: Population and event type distribution in GSR chains.

Table 3.1: Statistical properties of GSR and Non-GSR chains.

<table>
<thead>
<tr>
<th>Category</th>
<th>Avg-Time-Lag (hour)</th>
<th>% of Twitter starts</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSR Chains</td>
<td>10.95</td>
<td>40%</td>
</tr>
<tr>
<td>Non-GSR Chains</td>
<td>5.26</td>
<td>73%</td>
</tr>
</tbody>
</table>
Comparison of GSR with Non-GSR chains

Statistical properties of GSR and Non-GSR chains are depicted in Table 3.1. Here the column “% of Twitter Starts” refers to the percentage of chains where the interaction pattern starts with a information flow from Twitter to news i.e, chains with the qualitative encoding starting with “T” or “B”. Avg-Time-Lag refers to the average time difference between between any two consecutive news reports in a story chain. Some interesting facts from this table are: (1) the average time lag between two consecutive news reports in story chains is less than half a day. This is mainly because the period for which our data spans includes events such as the major soccer tournaments of the Confederations cup and mass protests known as the Brazilian Spring. (2) A lot of protest events, specifically approximately 40%, have Twitter starts indicating the presence of precursory signals in Twitter. Table 3.2 shows the breakup of the GSR chains in terms of event types and population which show that Govt. related, General Population and Other (Religious, Cultural, Social) events have a significant % of Twitter starts.

Difference in interaction pattern for GSR chains

We cluster the GSR chains into $K = 5$ clusters based on both qualitative and quantitative encoding(s) as described in Section 3.2.4. As detailed in the previous section, each GSR chain can be attached to a set of event types and populations. Fig. 3.6 gives the distributions of event type and populations for each of the 5 clusters of GSR chains. Both the clustering using the Jaro-Winkler distance metric on the qualitative encoding and the DTW metric on the quantitative encoding yield similar results and thus only one of them is reported here.

The variability in different population distributions across clusters is clearly evident. For instance, Cluster 0 has a greater share of story chains with population(s) such as “education”, “ethnic” and “labor”. Within the encoded sequences of this cluster, the most common sub-sequence at the beginning of the interaction pattern is (“NBNBT”) and (“E”). This indicates that at the beginning of such story chains, there is a news report coming out first and then Twitter discussion starts. Also, the alternative sub-patterns such as “NB” indicate that the discussion in Twitter involving these populations are not consistently active. In this cluster, chains encoded as “E”
(a) Population Distribution of Clusters (K-Medoids)  

(b) Event Type Distribution of Clusters (K-Medoids).

Figure 3.6: Stacked bar plots showing difference in event type and population distributions for different interaction pattern cluster. Cluster 4 patterns are mainly found in “other” related events and never in “energy”. Similarly, “medical” population type can only be seen in events having interaction patterns from clusters 0 and 3. Also events with “labor” as population fall primarily in cluster 0.

(with insignificant activity in Twitter) have population distribution over “agricultural”, “legal” and “business”. Clusters 2 and 3 are predominantly from “general population” with some share of “medical” and their encoded patterns look like “TBE”, “TNT”, “BNT” etc., which implies that for such clusters in general the Twitter is the first to break the story.

Regarding event types, there is diversity in terms of the distribution of “energy”, “govt. related” and “wages” related protests. Cluster 0 has a higher percentage of “wages” and “government” related protests. This cluster, in terms of population, consists of “education”, “ethnic”, and “labor”. Cluster 2 and 3 exhibit different event types. Cluster 3 has more story chains about “energy” and cluster 2 involves “gov” and “others”. Both of these two clusters have general population related patterns and more starters with Twitter. Cluster 4 has more proportion in “other” where also some proportion of “general population” events are found.

**Topic Variability in Interaction Patterns**

In order to understand the generic differences in topics exhibited by different clusters, we applied topic modeling to the documents in our dataset and used them to calculate topic distributions for each cluster as described in Section 3.2.5. For this set of experiments, we included both GSR and
Table 3.3: Important words in topics inferred by LDA.

<table>
<thead>
<tr>
<th>Topics</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>0. Economy</td>
<td>economia, brasil, milhao, bilhao, banco, mes, produto</td>
</tr>
<tr>
<td>1. Others</td>
<td>paulo, brasil, noticia, zap, portal, short, url, urllonga, anunciar</td>
</tr>
<tr>
<td>2. Government</td>
<td>governo, paulo, publico, projeto, presidente, dever, direito</td>
</tr>
<tr>
<td>3. Local Event</td>
<td>falar, saber, ceara, querer, mundo, nacional, clube, passar, fortaleza</td>
</tr>
<tr>
<td>4. Protest</td>
<td>manifestante, protesto, rio, feira, pessoa, avenida, polica</td>
</tr>
<tr>
<td>5. Entertainment</td>
<td>brasil, paulo, cultura, show, mostrar, cinema, gaga, filme</td>
</tr>
<tr>
<td>6. Internet</td>
<td>mail, cadastrar, login, senha, cpf, querer, abaixo, guardar</td>
</tr>
<tr>
<td>7. Business</td>
<td>empresa, podar, energia, mercado, setor, central, industria, servico</td>
</tr>
<tr>
<td>8. Crime</td>
<td>matar, pessoa, morrer, atingir, regiao, noticia, morte, violencia, vitima</td>
</tr>
<tr>
<td>9. Medical</td>
<td>rio, papa, janeiro, medicina, indio, maracana, francisco, hospital</td>
</tr>
<tr>
<td>10. International</td>
<td>internacional, america, eleicao, brasil, unidos, pais</td>
</tr>
<tr>
<td>11. Judicial</td>
<td>partido, presidente, dilma, federal, paulo, ministro</td>
</tr>
<tr>
<td>12. Advertisement</td>
<td>publicidade, brasil, rio, jornal, tolipan, helois, cultura</td>
</tr>
<tr>
<td>13. Transportation</td>
<td>paulo, polica, onibus, policial, ataque, capital, veiculo</td>
</tr>
<tr>
<td>14. Sports</td>
<td>copa, futebol, jogo, esporte, brasileiro, selecao, paulo</td>
</tr>
<tr>
<td>15. Geography</td>
<td>rio, chuva, cidade, casa, regiao, pessoa, janeiro,</td>
</tr>
<tr>
<td>16. Police</td>
<td>policia, policial, militar, matar, crime, caso, preso</td>
</tr>
<tr>
<td>17. Disaster</td>
<td>boate, maria, santo, incendio, pessoa, sul, tragedia, kiss</td>
</tr>
<tr>
<td>18. Technology</td>
<td>tecnologia, ciencia, sol, vinho, anna, ramalhar, ambiental, cultura</td>
</tr>
</tbody>
</table>

Table 3.4: Top topics for clusters of interaction patterns.

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Frequent Sub-patterns</th>
<th>Top Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0</td>
<td>“NBNBTNTN”, “NTNTN”</td>
<td>Local Events</td>
</tr>
<tr>
<td>C1</td>
<td>“NT”, “NTNT”</td>
<td>Local Events</td>
</tr>
<tr>
<td>C2</td>
<td>“TNT”</td>
<td>Local Events, Ads, Technology</td>
</tr>
<tr>
<td>C3</td>
<td>“T”, “TB”</td>
<td>Others, Protest, Sports</td>
</tr>
<tr>
<td>C4</td>
<td>“TNENT”, “TEB”</td>
<td>Protest, Government, Entertainment</td>
</tr>
</tbody>
</table>
Figure 3.7: Topic distributions of each interaction pattern cluster. The X-axis labels refer to topic number as given in Table 3.3. Also for each cluster a pattern cloud of the most significant interaction patterns found in the cluster is shown. For example, Cluster 3 has high proportion of topic 1 and mainly has story chains where information flow is mainly from Twitter to News.

Non-GSR chains. Table 3.3 shows the results from LDA. The topic labels were assigned manually by our domain expert. Fig 3.7 (next page) gives a general description of distributions of 20 topics over 5 clusters. Referring to Table 3.4 to get an idea of what each topic talks about, we can see that the distribution differences across the clusters in Fig 3.7 can be intuitively explained.

The discussions related to local events (natural and man-made) peaked in news before Twitter, though there could have been a few tweets about the event earlier than news. This can be seen by the appearance of $N$ in the frequent sub-patterns in such story chains. This inference is probably local to the type of events (fire accident) that happened during the period of analysis and of the geographic region (Brazil) studied in our dataset. Moreover, we see reciprocity between news and Twitter via subsequent alternating appearances of $B$, $N$ and $T$ states. In contrast, stories related to Sports, Protests and Advertisements tend to become hot trends earlier by Twitter followed by back and forth interactions with the traditional news media (as seen by the frequent sub-patterns for the chains in clusters 2, 3 and 4).

**Which one is the main influencer?**

We define the influence weight of a story chain as the average of the difference of pre- and post-influence weights, $\sum_{i} (W_{pre}^{i} - W_{post}^{i})/n$, where the summation is over $n$, the number of articles in a chain. This influence weight is used to identify the main influencer for a story chain i.e., which direction the information flow is predominant over its course. If the influence weight is positive,
Table 3.5: Story Chains (SC) with Interaction Patterns (IP) and Main Influencer (MI).

<table>
<thead>
<tr>
<th>ID</th>
<th>IP</th>
<th>Influence Weight</th>
<th>MI</th>
<th>Story Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC1</td>
<td>TT</td>
<td>0.514</td>
<td>Twitter</td>
<td>“Marco Feliciano enfrenta protesto na porta de igreja”</td>
</tr>
<tr>
<td>SC2</td>
<td>TN</td>
<td>0.48</td>
<td>Twitter</td>
<td>“25% Teachers are on strike. Government denies.”</td>
</tr>
<tr>
<td>SC3</td>
<td>NNNBNTBN</td>
<td>-0.422</td>
<td>News</td>
<td>“Fire in Kiss Nightclub in Santa Maria”</td>
</tr>
<tr>
<td>SC4</td>
<td>NBNNTN</td>
<td>-0.405</td>
<td>News</td>
<td>“Governor Genro decrees official mourning”</td>
</tr>
<tr>
<td>SC5</td>
<td>TTTNN</td>
<td>5.0e-05</td>
<td>Both</td>
<td>“After 8 years, Nadal back to Brazil with high investment and large team”</td>
</tr>
<tr>
<td>SC6</td>
<td>NNTTNNTN</td>
<td>-1.7e-04</td>
<td>Both</td>
<td>“Nissan sells more than 100 thousand first”</td>
</tr>
</tbody>
</table>

A larger absolute value implies Tweets are more active in this story chain and the flow of information is mostly from Twitter to news. If the weight is negative, the larger absolute value indicates news is mostly earlier than Tweets in reporting the sub-events within a story chain. Influence weight close to zero indicates significant reciprocal interaction between news and Twitter over the course of the story chain.

Table 3.5 (shown on next page) lists the top most chains with the highest positive, and negative influence weights as well as chains with approximately zero weights. We also present the corresponding interaction patterns, and a brief description of the story chain. For instance, story chain SC1, whose main influencer is Twitter talks about a student organized protest at the door of a church against the appointment of pastor Marco Feliciano at the presidency of Human rights of the federal chamber in Sao Paulo, Brazil. Story SC5 talks about the arrival of tennis player Rafael Nadal for Brazil’s only ATP tournament after 8 long years causing much hype in both traditional and social media. Overall, we can see that news domains track events related to politics and natural disasters, much earlier than Twitter, whereas Twitter is quicker in capturing information about social events and famous figures.
3.5 Summary

In this chapter, we presented a new framework for discovering the direction of information flow over time across two heterogeneous information media - news and Twitter. This lets us uncover the interaction patterns over stories consisting of chronologically chained news articles. We tested the proposed interaction pattern framework on real data from Brazil and we found that both Twitter and traditional News media have various influences on different topics. Twitter as a social network platform serves as a fast way to draw attention from public for many social events such as sports news whereas news media is quicker to report events regarding political, economical and business issues. Certain topics were found to have similar influence from both Twitter and News media.
Chapter 4

Multi-Instance Learning for Mining Precursors

Open source data (e.g., social media and news feeds) have been proven to serve as surrogates in forecasting a broad class of events such as political events and disease outbreaks [46]. From the perspective of human analysts and policy makers, forecasting algorithms must not only make accurate predictions but must also provide supporting evidence, e.g., the causal factors related to the event of interest. In this chapter, we present a novel multiple instance learning based approach that jointly tackles the problem of identifying evidence-based precursors and forecasts events into the future. Specifically, given a collection of streaming news articles from multiple sources we develop a nested multiple instance learning approach to forecast significant societal events such as protests. Using data from three countries in Latin America, we demonstrate how our approach is able to consistently identify news articles considered as precursors for protests. Our empirical evaluation demonstrates the strengths of our proposed approach in filtering candidate precursors, in forecasting the occurrence of events with a lead time advantage and in accurately predicting the characteristics of civil unrest events.
4.1 Background

Forecasting societal uprisings such as civil unrest movements is an important and challenging problem. Open source data (e.g., social media and news feeds) have been proven to serve as surrogates in forecasting a broad class of events, e.g., disease outbreaks [1], election outcomes [44, 61], stock market movements [7] and protests [46]. While many of these works focus on predictive performance, there is a critical need to develop methods that also yield insight by identifying precursors to events of interest.

This chapter focuses on the problem of identifying precursors (evidence) for forecasting significant societal events, specifically protests. Modeling and identifying the precursors for a given protest is useful for human analysts and policy makers as it discerns the underlying reasons behind the civil unrest movement. In particular, the objective is to study and forecast protests across different cities in three Latin American countries (Argentina, Brazil and Mexico). 6000 news outlets are tracked daily across these countries with the goal of forecasting protest occurrences with at least one day of lead time. From the news feeds, we also aim to identify the specific news articles that can be considered as precursors for the target event.

Figure 4.1 shows an example of precursors identified by our model. On the right of the timeline is a news report about a protest event in Argentina. The connected dots denotes the generated

Figure 4.1: Precursor story line for a protest event in Argentina. The x-axis is the timeline. The dots above with numbers are the probabilities for each day that the model generated for the target event. Each precursor document is titled in the timeline.
probabilities of a protest event over the days leading up to this protest. From this example, we find that within 10 days before the event, there are multiple precursor events identified as highly probable leading indicators of a protest. Most significant societal events are a consequence of several factors that affect different entities within communities and their relationships with each other (or the government) over time. In this specific example, the leading precursor was an article commenting on standards of living in Argentina and rising poverty levels. The International Court of Justice also delivered a verdict on the debt crisis. All these factors led to the final protest involving the general population across the country demanding better work opportunities.

We formulate the precursor identification and forecasting problem in a novel multiple instance learning algorithm (MIL) setting. Multiple instance learning algorithms [3, 72] are a class of supervised learning techniques that accept labels for groups of instances, but where labels for individual instances are not available. In our formulation, instances denote news articles and while class labels are not associated with individual news articles, a group of news articles is attached with a label (indicating the occurrence of a protest). We further extend the standard MIL formulation by introducing a nested structure, wherein we group news articles published in a given day at the first level and then group the collection of individual days at the second level. This nested MIL approach allows for modeling the sequential constraints between the news articles (grouped by days) published on different days and also provides a probabilistic estimate for every news article and the collection of news articles. This estimate is significant because it indicates for a given news article the probability of it signaling a protest event. Recall that in our datasets we do not have any training labels to associate a protest per news article.

The main contributions of this study are summarized as follows:

1. **A novel nested framework of multi-instance learning for event forecasting and precursor mining.** We formulate event forecasting and precursor mining for multiple cities in a country as a multi-instance learning problem with a nested structure. By estimating a prediction score for each instance in the history data, we automatically detect significant precursors for different events.

2. **Harness temporal constraints in multi-instance learning.** We explore different penalty function and regularizations where we employ the temporal information in our dataset under
assumption that most events of interest are follow-up reports of other events that happened before, and most planned events are developing over time.

3. Modeling for various event categories in multiple geo-locations. We extend the nested MIL formulation for general purpose multi-class classification to determine necessary attributes of events in terms of their underlying population.

4. Application and evaluation with comprehensive experiments. We evaluate the proposed methods using news data collected from July 2012 to December 2014 in three countries of Latin America: Argentina, Mexico, and Brazil. For comparison, we implement other multi-instance algorithms, and validate the effectiveness and efficiency of the proposed approach. We also perform qualitative and quantitative analysis on the precursors inferred by our model.

4.2 Problem Formulation

Given a collection of streaming media sources (e.g., news feeds, blogs and social network streams), the objective of our study is to develop a machine learning approach to forecast the occurrence of an event of interest in the near future. Specifically, we focus on forecasting protests or civil unrest movements in Latin America from a daily collection of published news articles. Besides forecasting the protest, we aim to identify the specific news articles from the streaming news outlets that can be considered as supporting evidence for further introspection by an intelligence analyst. We refer to these identified articles as precursors for a specific protest.

Figure 4.2 provides an overview of our proposed approach and problem formulation. Here, we show groups of news articles collected daily, five days prior to the specific protest event (being forecast). Within our proposed MIL-based formulation, each news article is an individual instance, the collection of news articles published on a given day is a bag, and the ordered collection of bags (days) is denoted as a super-bag (explained in detail later). For this study, each individual news article is represented by a distributed representation for text derived using a framework such as text embedding [33]. Figure 4.2 shows that for certain days within the collection we attempt to identify news articles (highlighted) that are considered as precursors from the entire collection of input news articles used for forecasting the occurrence of a specific target.
Table 4.1: Notations used in the nMIL model.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{S}$</td>
<td>a set of $n$ “super bags” and their labels, $\mathcal{S} = {\mathcal{S}, Y}$, in our dataset</td>
</tr>
<tr>
<td>$N$</td>
<td>number of superbags in our dataset</td>
</tr>
<tr>
<td>$\mathbb{S}$</td>
<td>a “super bag” which is also an ordered set of $t$ “bags”, $\mathbb{S} = [\mathbf{X}_i], i \in {1, ..., t}$</td>
</tr>
<tr>
<td>$\mathbf{X}_i$</td>
<td>a bag which is a set of instances, $\mathbf{X}<em>i = {x</em>{ij}}, j \in {1, ..., n_i}$ with $n_i=</td>
</tr>
<tr>
<td>$x_{ij}$</td>
<td>the $j$-th instance in set $\mathbf{X}_i$, a V-dimension vector, $\mathbb{R}^{V \times 1}$</td>
</tr>
<tr>
<td>$Y$</td>
<td>event label $\in {-1, +1}$ of super bag $\mathcal{S}$</td>
</tr>
<tr>
<td>$P$</td>
<td>estimated probability $\in [0, 1]$ for a super bag</td>
</tr>
<tr>
<td>$P_i$</td>
<td>the probability $\in [0, 1]$ of bag $i$ in a super bag to be positive</td>
</tr>
<tr>
<td>$p_{ij}$</td>
<td>the probability of an instance $x_{ij}$ in bag $\mathcal{X}_i$ in a super bag to be positive w.r.t an event</td>
</tr>
<tr>
<td>$O$</td>
<td>a set of precursor documents for an event</td>
</tr>
<tr>
<td>$C$</td>
<td>multi-class label of super bag $C \in {1, 2, ..., K}$</td>
</tr>
<tr>
<td>$w$</td>
<td>model parameter in the nMIL model, a V-dimension vector $\mathbb{R}^{V \times 1}$.</td>
</tr>
</tbody>
</table>
4.2.1 Formal Definition and Notation

For a given protest event $e$ occurring on day $t + l$, we assume that for each day before the event we are tracking a multitude of news sources. The occurrence of the protest event at time $t + l$ is denoted by $Y_{t+l} \in \{-1, +1\}$ where 1 denotes a protest and $-1$, otherwise.

**Definition 1** A bag is a collection of events reported on a given day $i$ by $X_i = \{x_{i,1}, \ldots, x_{i,n_i}\}$, where $x_{i,j}$ is the $j$-th event represented by a vector.

**Definition 2** A positive super-bag is the ordered collection of events before a protest event up to day $t$: $S_{1:t} = \{X_1, \ldots, X_t\}$.

**Definition 3** The forecasting problem can be formulated as learning a mathematical function $f(S_{1:t}) \rightarrow Y_{t+l}$ that maps the input, an ordered collection of news articles extracted per day to a protest indicator $l$ days in the future from the day $t$.

**Definition 4** We aim to estimate the probability for each event on any given day that signifies the occurrence of a given protest. For an input $x_{ij}$, we denote this estimated probability value by $p_{ij}$. The precursor discovery is to identify the events in a super-bag as the ones with:

$$p_{ij} \geq \tau$$

We represent this precursor set of documents as a subset of the original super-bag, given by $O = \{x_{ij} \in S_{1:t} | p_{ij} > \tau\}$.

As a secondary objective, we aim to forecast the occurrence of an event with a long lead time i.e., large values of $l$. Table 4.1 captures the notation and definitions used in this study.
4.3 Methodology

We cast our event forecasting problem into a classification problem and resort to multiple instance learning with deep features to train our model. Mining precursors for significant events will automatically employ the knowledge discovered in the classification.

We first provide our intuition behind formulating the precursor discovery and forecasting problem within a novel extension of multiple instance learning algorithm. Parallel to the standard multiple instance learning algorithms we have a group of news articles (bags) with labels available only for the entire bag (i.e., leading to a protest); and one of the objectives is to train a classifier to predict the bag-level label. In addition to predicting the group-level labels, we also care about predicting the labels for individual news articles (instances) since they signify the precursor. Various MIL formulations extend the basic definition with a similar motivation, i.e., to estimate the key instances within a bag or provide instance-level labels. However, our problem setting has a two-level grouping structure with sequential constraints, i.e., we capture news articles per day (bags) and group the days to form a super-bag with labels only available at the super-bag level. As such, we propose a nested multiple instance learning formulation for predicting the super bag level labels (forecast) and then estimate the bag-level and instance-level probabilities for identifying association of the bag and instance with the event, respectively. We developed various extensions of our proposed approach to tie the different sequential and group constraints.
4.3.1 Nested MIL model (nMIL)

We model the instance level probability estimates $p_{ij}$ for a news article $j$ on day $i$ to associate with a targeted event $e$ with a logistic function. These probability estimates indicate how related the specific instance is to the target event, $e$. Higher the probability value, the more related the document is to the target event and most probably represents a precursor that contains information about causes of the target event.

$$p_{ij} = \sigma(w^T x_{ij}) = \frac{1}{1 + e^{-w^T x_{ij}}}.$$  \hfill (4.1)

Here, $w$ denotes the learned weight vector for our model. The probability for a day (or bag) is then modeled as the average of probability estimates of all instances in a day [29]. Hence, for each bag:

$$P_i = A(X_i, w) = \frac{1}{n_i} \sum_{j} p_{ij},$$ \hfill (4.2)

where $A$ is an aggregation function.

We then model the probability of a super-bag $S$ (associated with an event $e$) being positive as the average of the probability of all $t$ bags within the super bag to be positive (related to the target event). Thus:

$$P = A(S, w) = \frac{1}{t} \sum_{i} P_i$$ \hfill (4.3)

For a given super bag $S$, as all the $t$ bags within it are temporally ordered, the probability estimates for a given bag (day) is assumed to be similar to its immediate predecessor. This consistency in consecutive bag probabilities is modeled by minimizing the following cross-bag cost as below:

$$g(X_i, X_{i-1}) = (P_i - P_{i-1})^2$$ \hfill (4.4)

Finally, given a set of true labels $Y$ for the super bags, we can train our model by minimizing the following cost function w.r.t to $w$:
\[
J(w) = \beta \sum_{S \in \mathcal{S}} f(S, Y, w) + \frac{1}{n} \sum_{S \in \mathcal{S}} \frac{1}{t} \sum_{i=1}^{t} g(X_i, X_{i-1}, w) + 1 \sum_{S \in \mathcal{S}} \sum_{X_i, X_{i-1} \in S} X_i \cdot X_{i-1} \cdot \mathbb{1} \sum_{i=1}^{n} \sum_{j=1}^{n} h(x_{ij}, w) + \lambda R(w)
\] (4.5)

Here,

- \( f(S, Y, w) = -I(Y = 1)\log P - I(Y = -1)(\log(1 - P)) \) is the negative log-likelihood function that penalizes the difference between prediction and the true label for super bag \( S \) where \( I(\cdot) \) is the indicator function.

- \( g(X_i, X_{i-1}, w) \) is the cross-bag cost defined in Equation 4.4.

- \( h(x_{ij}, w) = \max(0, m_0 - \text{sgn}(p_{ij} - p_0)w^T x_{ij}) \) represents the instance level cost. Here, \( \text{sgn} \) is the sign function; \( m_0 \) is a crucial margin parameter used to separate the positive and negative instances from the hyper line in the feature space; \( p_0 \) is a threshold parameter to determine positiveness of instance.

- \( R(w) \) is the regularization function.

- \( \beta, \lambda \) are constants that control the trade-offs between the loss function and regularization function.

**Cross-bag Similarity (nMIL\(^\Delta\))**

The cross-bag similarity \( g(\cdot, \cdot) \) in the above equation does not allow for sudden changes in the day-level probabilities caused due to newer events happening on the current day. We update the cost function across days (bags) (Equation 4.4) as follows:

\[
g(X_i, X_{i-1}) = \Delta(X_i, X_{i-1})(P_i - P_{i-1})^2
\] (4.6)

The objective function above allows for label information to spread over the manifold in the feature-space. As such, we compute \( \Delta(\cdot, \cdot) \) as the pairwise cosine similarity between the news articles in \( X_i \) and \( X_{i-1} \). Since we do not have ground truth labels for the bag level (day) we make this consistency assumption that estimated probabilities for consecutive days should be similar if the
news articles have similarity in the feature space as well. This model is referred by nMILΔ and allows for sudden changes in how events unfold.

### 4.3.2 Sequential Model (nMILΩ)

The basic nMIL models assume that there exists a single weight vector across all the days (bags) within a super bag. To model the sequential characteristics of the articles published across consecutive days, we extend this formulation by learning individual weight vectors for each of the historical days. Assuming $t$ days within a super bag $S$ we learn a weight vector for each individual day represented as $\Omega = [w_1, \ldots, w_t]$; where $w_i$ is the weight vector learned for day $i$. In this setting, the individual weight vectors are still learned together in a joint fashion akin to multi-task learning approaches [9].

However, the probability of a news article $j$ on day $i$ will be given by $p_{ij} = \sigma(w_i^T x_{ij})$. This formulation is called nMILΩ and given by:

$$J(\Omega) = \frac{\beta}{n} \sum_{S \in S} f(S, \Omega, Y) + \frac{1}{n} \sum_{S \in S_i} \frac{1}{t} \sum_{i=1}^{t} g(X_i, X_{i-1}, w_i)$$

$$+ \frac{1}{n} \sum_{S \in S} \frac{1}{n_i} \sum_{j=1}^{n_i} h(x_{ij}, w_i) + \lambda R(\Omega)$$

Just like the multi-task learning algorithms, the regularization term $R(\Omega)$ can be modified to capture the various relationship-based constraints. However, in this study we ignore these specialized approaches focusing only on the MIL paradigm.

### 4.3.3 Multiclass Classification

We also extend our developed nMIL formulations to solve general purpose multiclass classification problems rather than binary classification problems.
Within our domain, each labeled event is manually attached with event population. Event population indicates the size/community of people who participated in the protest event.

For multiclass classification problems, we train one-versus-rest classifiers for each of the classes learning a separate weight vector per class. When classifying a super bag to a specific event population we first forecast the binary protest indicator label for a super bag. Next, we apply the multi-class classification only on the predicted positive examples.

### 4.3.4 Optimization

We perform online stochastic gradient decent optimization to solve our cost function and test our model on new data to predict super bag label. For every iteration in our algorithm, we randomly choose a super-bag \((S, Y)\) from the training dataset \(S\) by picking an index \(r \in \{1, \ldots, n\}\) using a standard uniform distribution. Then we optimize an approximation based on the sampled super-bag by:

\[
J(w; S) = \beta f + \frac{1}{t} \sum_{i} g_{i} + \frac{1}{t} \sum_{i} \frac{1}{n_{i}} \sum_{j} h_{ij} + \lambda R(w)
\]  

(4.8)

The gradient of the approximate function is given by:

\[
\nabla J(w) = \frac{\partial J(w; S)}{\partial w} = \lambda w
\]

(4.9)
where \( v = i - 1, o_{ij} = I(\text{sgn}(p_{ij} - p_0)wx_{ij} < m_0) \). We update the weight vector using a varied learning rate and \( w' = w - \eta \nabla(w) \) using mini-batch stochastic gradient descent where \( \eta \) is the learning rate at current iteration.

### 4.3.5 Precursor discovery using nMIL

In the nMIL model, each super-bag consists of an ordered set of bags and each bag represents the documents in one day in the city for which we are forecasting a protest event. We present in Algorithm 1 the steps to identify news articles as precursors based on their estimated probability given by \( p_{ij} > \tau \).

**Algorithm 2 Precursor Discovery in nMIL**

1: procedure PD-nMIL
2: Input: \( S = \{(S_r, Y_r)\}_{r \in \mathbb{N}^+}, \mathcal{M} \)
3: Output: \( \{(O_r, Y_r)\}_{r \in \mathbb{N}^+} \)
4: for super bag \( (S_r, Y_r) \) do
5: \( O_r = [] \)
6: for \( t = 1, 2, ..., h \) (history days) do
7: \( y_t = [] \)
8: for \( x_{tm} \in \mathcal{X}_t \) do
9: \( \hat{y}_{tm} = \sigma(\hat{w}x_{tm}) \)
10: if \( \hat{y}_{tm} > \tau \) then
11: \( y_t \leftarrow (m, \hat{y}_{tm}) \)
12: sort\( (y_t) \) by \( \hat{y}_{tm} \) in descending order
13: \( O_r \leftarrow m \) where \( m \) in top\( (y_t) \)
14: return \( \{(O_r, Y_r)\}_{r \in \mathbb{N}^+} \)
4.4 Experiments

4.4.1 Datasets

The experimental evaluation was performed on news documents collected from around 6000 news agencies between July 2012 to December 2014 across three countries in South America, viz. Argentina, Brazil, and Mexico. For Argentina and Mexico, the input news articles were primarily in Spanish and for Brazil, the news articles were in Portuguese.

The ground truth information about protest events, called the gold standard report (GSR) is exclusively provided by MITRE [46]. The GSR is a manually created list of civil unrest events that happened during the period 2012-2014. A labeled GSR event provides information about the geographical location at the city level, date, type and population of a civil unrest news report extracted from the most influential newspaper outlets within the country of interest. These GSR reports are the target events that are used for validation of our forecasting algorithm. We have no ground truth available for verifying the validity of the precursors.

Argentina: We collected data for Argentina from newspaper outlets including Clarin and Lanacion for the period of July 2010 to December 2014. There are multiple protest events in Argentina during this period. For instance, people protested against the government and utility/electricity-providing companies because of heatwaves in Dec. 2013.

Brazil: For Brazil, we obtained data from news agencies including the three leading news agencies in Brazil; O Globo, Estadao, and Jornal do Brasil from November 2012 to September 2013. During this period Brazil faced several mass public demonstrations across several Brazilian cities stemming from a variety of issues ranging from transportation costs, government corruption, and police brutality. These mass protests were initiated due to a local entity advocating for free public transportation. This period had an unusually high social media activity and news coverage and is also known as the “Brazilian Spring” 1.

Mexico: For Mexico, we tracked news agencies including the top outlets: Jornada, Reforma, Milenio from January 2013 to December 2014. Over 619 days, we noticed 71 news articles per day

1http://abcnews.go.com/ABC_Univision/brazilian-spring-explainer/story?id=19472387
on average. There were more than 2000 protest events in this two-year period with major unrest movements in 2013 led by teachers and students demanding education reforms by protesting against the government.

4.4.2 Experimental Protocol

The GSR signifies the occurrence of a protest event on a given day at a specific location. To evaluate the MIL-based forecasting and precursor discovery algorithms, for each protest event we extract all the published news articles for up to 10 days before the occurrence of the specific event. This ordered collection of per-day news documents up to the protest day are considered as positive super bags. For negative samples, we identify consecutive sets of five days within our studied time periods for the different countries when no protest was reported by the GSR. The ordered collection of per-day news documents not leading to a protest are considered as negative super bags for the nMIL approach. For any news article (i.e., an individual instance) within a positive/negative super-bag we have no label (or ground truth). As part of the precursor discovery algorithm, we estimate a probability for an individual instance to signal a protest (by showing evidence). It is important to note that the GSR linked news article for a protest is never used for training purposes. Having identified the positive and negative samples, we split our datasets into training and testing partitions and perform 3-fold cross-validation. A single run of the model on a machine with 4 cores and 16 GB memory takes about 250 seconds.

We study the performance of forecasting models with varying lead time and varying historical days. Lead time ($l$) indicates the number of days in advance the model makes predictions and historical days ($h$) denotes the number of days over which the news articles are extracted as input to the prediction algorithms. As an example, if $l$ is set to 1, then the model forecasts if a protest event is planned for the next day. Setting the historical days, $h$, to 5 denotes that we use news from five days before the current day to make the forecast. We varied $l$ from 1 to 5 and $h$ from 1 to 10 and trained 50 different models for the different approaches to study the characteristics of the developed approaches with varying lead time and historical days.

For event forecasting, we evaluate the performance by standard metrics including precision, recall, accuracy and F1-measure.
4.4.3 Comparative Approaches

We compare the proposed nMIL models to the following approaches:

- **SVM**: We use the standard support vector machine formulation [12] by collapsing the nested grouping structure and assigning the same label for each news article as its super-bag (for training). During the prediction phase, the SVM yields the final super-bag prediction (forecast) by averaging the predicted label obtained for each of the instances.

- **MI-SVM [3]**: The MI-SVM model extends the notion of a margin from individual patterns to bags. Notice that for a positive bag the margin is defined by the margin of the “most positive” instance, while the margin of a negative bag is defined by the “least negative” instance. In our case, we collapse the news articles from the different historical days into one bag and apply this standard MIL formulation.

- **Relaxed-MIL (rMIL$^{nor}$) [65]**: Similar to the MI-SVM baseline, we collapse the news articles into one bag. However, unlike the MI-SVM formulation the rMIL$^{nor}$ model can provide a probabilistic estimate for a given document within a bag to be positive or negative.

- **Modified Relaxed-MIL (rMIL$^{avg}$)**: This approach is similar to the rMIL$^{nor}$, except we compute the probability of a bag being positive by taking average of estimate of each instance in the bag rather than using the Noisy-OR model discussed above.

- **GICF [29]**: This model optimizes a cost function which parameterized the whole-part relationship between groups and instances and pushes similar items across different groups to have similar labels.

4.4.4 Feature Description

In practice, finding good feature representations to model the news articles is not a trivial problem. Traditionally the bag-of-words representation allows for easy interpretation but also requires pre-processing and feature selection.

Several researchers have developed efficient and effective neural network representations for language models [5, 38, 40]. Specifically, we learn deep features for documents by taking advantage of the
Table 4.2: Event forecasting performance comparison based Accuracy (Acc) and F-1 score w.r.t to state-of-the-art methods. The proposed nMIL, nMIL$^\Delta$, nMIL$^\Omega$ method outperform state-of-the-art methods across the three countries.

<table>
<thead>
<tr>
<th></th>
<th>Argentina</th>
<th></th>
<th>Brazil</th>
<th></th>
<th>Mexico</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>F-1</td>
<td>Acc</td>
<td>F-1</td>
<td>Acc</td>
<td>F-1</td>
</tr>
<tr>
<td>SVM</td>
<td>0.611(±0.034)</td>
<td>0.406(±0.072)</td>
<td>0.693(±0.040)</td>
<td>0.598(±0.067)</td>
<td>0.844(±0.062)</td>
<td>0.814(±0.091)</td>
</tr>
<tr>
<td>MI-SVM</td>
<td>0.676(±0.026)</td>
<td>0.659(±0.036)</td>
<td>0.693(±0.040)</td>
<td>0.503(±0.087)</td>
<td>0.880(±0.025)</td>
<td>0.853(±0.040)</td>
</tr>
<tr>
<td>rMILavg</td>
<td>0.330(±0.040)</td>
<td>0.411(±0.092)</td>
<td>0.505(±0.012)</td>
<td>0.661(±0.018)</td>
<td>0.499(±0.009)</td>
<td>0.655(±0.025)</td>
</tr>
<tr>
<td>GICF</td>
<td>0.589(±0.058)</td>
<td>0.624(±0.048)</td>
<td>0.650(±0.055)</td>
<td>0.649(±0.031)</td>
<td>0.770(±0.041)</td>
<td>0.703(±0.056)</td>
</tr>
<tr>
<td>nMIL</td>
<td>0.709(±0.036)</td>
<td>0.702(±0.047)</td>
<td>0.723(±0.039)</td>
<td>0.686(±0.055)</td>
<td>0.898(±0.031)</td>
<td>0.902(±0.030)</td>
</tr>
<tr>
<td>nMIL$^\Delta$</td>
<td>0.708(±0.039)</td>
<td>0.714(±0.034)</td>
<td>0.705(±0.048)</td>
<td>0.698(±0.045)</td>
<td>0.861(±0.014)</td>
<td>0.868(±0.014)</td>
</tr>
<tr>
<td>nMIL$^\Omega$</td>
<td>0.687(±0.038)</td>
<td>0.680(±0.045)</td>
<td>0.713(±0.028)</td>
<td>0.687(±0.038)</td>
<td>0.871(±0.013)</td>
<td>0.879(±0.014)</td>
</tr>
</tbody>
</table>

existing doc2vec model. For each document, we generate a 300 dimension vector for training with a contextual window size of 10 in an unsupervised version [33]. We compared the performance of deep features with traditional TF-IDF features but the results showed little difference. Thus, we only report the evaluation of models with deep features.

### 4.5 Results and Discussion

In this section, we evaluate the performance of the proposed models. First, we evaluate the effectiveness and efficiency of the methods on real data in comparison with baseline methods on multiple configurations of forecasting tasks. Then, we study and analyze the quality of precursors with respect to quantitative and qualitative measures. Multi-class forecasting evaluation is also provided for one of the countries. Finally, we perform a sensitivity analysis of performance regarding parameters in the proposed model.
Figure 4.3: Forecasting evaluation on 3 countries with respect to F1 score for SVM, rMIL, nMIL, and nMILΔ. X-axis is the number of historical days used in the training process. Y-axis shows the average F1 score of 10 runs of experiments.

4.5.1 How well does nMIL forecast protests?

Comparative Evaluation

Table 4.2 reports the prediction performance of the nMIL approach in comparison to other baseline approaches for the task of forecasting protests. Specifically, we use set $\beta = 3.0$, $\lambda = 0.05$, $m_0 = 0.5$ and $p_0 = 0.5$ ($\beta, \lambda$ chosen by sensitivity analysis, $m_0, p_0$ by default setting in hinge loss) and report the average accuracy and F1 score along with standard deviation for predicting protests across multiple runs of varying historical days with lead time set to 1. We observe that the nMIL approaches outperform the baseline approaches across all the three countries. The rMIL approach performs poorly because the the noisy-OR aggregation function associating the bag-level labels to instance-level labels forces most of the news articles within the positive bags to have probability values close to 1. However, given the large collection of news articles available per day only a subset of them will provide a signal/evidence for a protest. For Argentina, the nMIL and nMILΔ approaches outperformed the best baseline (MI-SVM), by 7% and 8% with respect the average F1 score, respectively.

Figure 4.3 shows the changes to F1 score for the proposed nMIL approach in comparison to SVM, MI-SVM, and rMILavg for different number of historical days that are used in training with lead time set to 2. We trained 10 different models that use different number of historical days respectively varying from 1 to 10. These results show that the methods that utilize the nested structure (nMIL, nMILΔ) within the multi-instance learning paradigm generally performed better than others.
Moreover, the proposed nMIL models performed well consistently across different countries with different number of history days.

**How early can nMIL forecast?**

In order to study the changes of performance with and without the nested structure, we show the F1 score with varying lead times and historical days from 1 to 5 for rMIL\textsuperscript{avg} and nMIL\Delta models in Table 4.3, respectively. We observe that with larger lead time (i.e., forecasting earlier than later), the nMIL model does not necessarily lose forecasting accuracy, but is sometimes even better. This can be explained by the fact that several times protests are planned a few days in advance and that civil unrest unfold as a series of actions taken by multiple participating entities over a sequence of days. As the lead time increases, F1 score for forecasting initially drops and then increases back. This behavior is also noted in prior work by Ramakrishnan et. al [46], which includes protest related data from these countries. In comparison to the nMIL model, the rMIL\textsuperscript{avg} approach, (which collapses the sequential structure encoded within the history of days) seems to perform inconsistently with increasing lead time.

### 4.5.2 Do the precursors tell a story?

**Quantitative Evaluation**  Figures 4.4a and 4.4b show the distribution of the estimated probabilities for instances within positive and negative super bags for Argentina and Mexico, respectively.

The instances within the negative super bags show lower probability estimates by the proposed model and the instances within the positive super bags show higher probability estimates. For Mexico, fewer instances within the positives are assigned high probabilities indicating strength of the proposed model to identify and rank the precursors.

Relative cosine similarity is computed as the pairwise normalized cosine similarity, scaled relative to each event.

Figures 4.5a and 4.5b show the average cosine similarity value for the precursor documents (probability estimate greater than 0.7) with the target GSR documents.
Table 4.3: F1-measure for rMIL^AV and nMIL^Δ models on Argentina, Brazil, and Mexico with historical days from 1 to 5 for leadtime 1 to 5.

<table>
<thead>
<tr>
<th>Country</th>
<th>Argentina</th>
<th>Brazil</th>
<th>Mexico</th>
</tr>
</thead>
<tbody>
<tr>
<td>History Days</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Leadtime 1</td>
<td>rMIL^AV</td>
<td>0.719</td>
<td>0.714</td>
</tr>
<tr>
<td></td>
<td>nMIL^Δ</td>
<td>0.745</td>
<td>0.735</td>
</tr>
<tr>
<td>Leadtime 2</td>
<td>rMIL^AV</td>
<td>0.659</td>
<td>0.624</td>
</tr>
<tr>
<td></td>
<td>nMIL^Δ</td>
<td>0.664</td>
<td>0.675</td>
</tr>
<tr>
<td>Leadtime 3</td>
<td>rMIL^AV</td>
<td>0.674</td>
<td>0.606</td>
</tr>
<tr>
<td></td>
<td>nMIL^Δ</td>
<td>0.649</td>
<td>0.669</td>
</tr>
<tr>
<td>Leadtime 4</td>
<td>rMIL^AV</td>
<td>0.656</td>
<td>0.558</td>
</tr>
<tr>
<td></td>
<td>nMIL^Δ</td>
<td>0.676</td>
<td>0.693</td>
</tr>
<tr>
<td>Leadtime 5</td>
<td>rMIL^AV</td>
<td>0.669</td>
<td>0.676</td>
</tr>
<tr>
<td></td>
<td>nMIL^Δ</td>
<td>0.626</td>
<td>0.676</td>
</tr>
</tbody>
</table>
Figure 4.4: Estimated probabilities for negative examples (purple) and positive examples (green) for Argentina and Mexico.

Figure 4.5: Mean of relative cosine values w.r.t target events in history days for Argentina and Mexico.
Figure 4.6: The figures on the left depict the distribution of relative cosine similarity for all documents (green line) and for precursor documents (blue line) with probability greater than 0.7. The figures on the right row depict the distribution of relative entity hit score for all documents (green line) and for precursor documents (blue line) with probability greater than 0.7.

For Argentina, we observe that on average, the documents on day 5 have the highest semantic similarity to the target event documents (GSR). The documents on day 3 and day 10 have lower similarity compared to the target event.

In order to investigate the relationship between the semantic similarity and the estimated probability by the proposed models, we compare the distribution of relative cosine similarity and relative entity hit score of the precursor documents with the target GSR documents with respect to bag of words features. Entity words in each news document are extracted by an enrichment tool for natural language processing. The relative entity hit score is calculated as the the intersection of entity set of precursor document and the target event divided by the relative minimum length of these two sets.

Figures 4.6a and 4.6b show the fitted Gaussian distribution of relative cosine similarities for all documents (green lines) and precursor documents (blue lines) for Argentina and Mexico, respectively. Figures 4.6c and 4.6d show the distribution of relative entity hit score for Argentina and Mexico, respectively. These distribution figures demonstrate that the proposed model assigns higher probability to news articles with higher semantic similarity to the GSR articles representing the protests events. These results show the strength of our proposed models in identifying the precursor articles.

**Case Studies** We present findings about the identified precursors based on the probability estimate by nMIL across three observed protests. In Figure 4.1, we present a protest event against
(a) A continuous police protest in Argentina against government for better salary. In the beginning, policemen at Cordoba were requesting for better salaries. Later on, police in Catamarca were involved in clashes with gendarmerie. Three days before the target event, the government sent out troops and more and more police joined for the same purpose. One day before the event, Buenos Aires state call for a strike.

(b) Protester in Mexico burned the congressional offices for justice for the missing students. In the beginning, students were marching for justice. Gradually, more communities such as artistic and policing community joined the event. Later on, children, youth, adults, students and teachers blocked traffic for protest.

Figure 4.7: Case Studies.
government in Argentina, and the selected precursors before its occurrence with their estimated probabilities. The titles of news reports as precursors are shown in the timeline.

In Figures 4.7a and 4.7b, we present story lines by precursors that were discovered for two different protest events in Argentina and Mexico, respectively. Figure 4.7a showcases the story line about a protest event in Argentina in December 2014. In this case, the police were protesting against government for better salaries. Before this event, clashes between police and gendarmerie (military policy) had occurred leading to the involvement of several policemen from different parts of the country. The text from news articles demonstrate the tense situation between the police and government in La Pampa, Argentina identified as precursors.

Figure 4.7b shows another story line of a continuous protest event in Mexico regarding the infamous case of 43 missing students\(^2\). The resulting outrage triggered constant protests which were identified by our proposed model. The figure shows a timeline of how the events turned violent leading up to the burning of congressional offices and depicts how different communities joined the movement.

4.5.3 Can nMIL forecast event populations?

We also evaluated the performance of our nMIL approaches for predicting the event populations by solving a multi-class classification problem. In Table 4.4 we depict the weighted-average F1 score for event populations (here, with categories such as Government, Wages, Energy, Others drawn from the GSR). Due to space limitations, we only depict the performance of weighted average F1 score on event population across 1 to 5 historical days with lead time of 1.

Table 4.4: Multi-Class F1-Measure for rMIL\(^{avg}\) and nMIL models on Argentina and Mexico with historical days from 1 to 5.

<table>
<thead>
<tr>
<th>History Days</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Average(Variance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rMIL(^{avg})</td>
<td>0.512</td>
<td>0.512</td>
<td>0.473</td>
<td>0.417</td>
<td>0.457</td>
<td>0.474(1e-3)</td>
</tr>
<tr>
<td>nMIL</td>
<td>0.523</td>
<td>0.552</td>
<td>0.515</td>
<td>0.485</td>
<td>0.537</td>
<td>0.524(7e-4)</td>
</tr>
<tr>
<td>Mexico</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rMIL(^{avg})</td>
<td>0.576</td>
<td>0.526</td>
<td>0.447</td>
<td>0.547</td>
<td>0.493</td>
<td>0.518(3e-3)</td>
</tr>
<tr>
<td>nMIL</td>
<td>0.570</td>
<td>0.583</td>
<td>0.560</td>
<td>0.615</td>
<td>0.545</td>
<td>0.575(7e-4)</td>
</tr>
</tbody>
</table>

\(^2\)https://en.wikipedia.org/wiki/2014_Iguala_mass_kidnapping
The proposed multi-class nMIL model outperforms the multi-class rMILavg model. On average, for event population, nMIL outperformed rMILavg by 10.5% and 10.6% for Argentina and Mexico, respectively.

4.5.4 How sensitive is nMIL to parameters?

![Image of graphs showing sensitivity analysis on β and λ.](a) β in nMIL (b) λ in nMIL

Figure 4.8: Sensitivity analysis on β and λ. x-axis represents the varying values for the parameter and y-axis denotes the test accuracy.

There are three main parameters in the proposed nMIL model, which are the regularization parameter λ, weight for super bag loss β and threshold for instance level hinge loss $m_0$. Figures 4.8a and 4.8b illustrate the performance of the proposed nMIL by varying β and λ, respectively. The test accuracy for different values of λ and β is relatively stable.

4.6 Summary

This chapter has presented a novel extension of the multi-instance learning framework for event forecasting and for identifying precursors for protest events. Most existing multi-instance approaches solve problems in object detection in images, drug activity prediction or identify sentimental sentences in text reviews. In contrast, we provide a novel application of MIL algorithms that require a two-level nested structure for event forecasting and precursor modeling. We have also studied the
strengths of our developed methods on open source news datasets from three Latin American countries. Through extensive evaluation and analysis, we illustrate the strong forecasting performance of the proposed methods with varying lead time and historical data. We also show qualitatively via several case studies, the richness of the identified precursors for different protests across different cities.
Chapter 5

Spatio-Temporal Precursor Learning for Event Forecasting

A global model presented in the last chapter is not always optimal. For instance, many societal events are considered as a system of inter-connected locations, where each location is recording a set of time-dependent observations. In order to detect event occurrence and automatically reconstruct the precursors and signals, it is essential to model relationships between the different locations w.r.t. how events evolve over time. However, existing methods for precursor discovery do not capture or exploit spatial and temporal correlations inherent in event occurrences. The absence of such modeling not only creates shortcomings in the quality of inference but also curtails interpretation by human analysts. Furthermore, forecasting is inhibited when training data is sparse. In this chapter, we develop a novel multi-task model with dynamic graph constraints within a multi-instance learning framework. Our model tackles the problem of scarce data distribution and reinforces co-occurring location-specific precursors with augmented representations. Through studies on civil unrest movements in numerous countries, we demonstrate the effectiveness of the proposed method for precursor discovery and event forecasting.
5.1 Background

While studying large scale societal events, policy makers and professionals aim to reconstruct precursors to such events to help understand causative attributes. Given a document reporting an event of interest (e.g., a civil protest), precursors are other documents published earlier than reported incidents or happenings and can be viewed as influential in the lead up to the protest. Such analysis is typically done painstakingly with the aid of subject matter experts, but new algorithmic tools [43] have recently emerged that support such precursor discovery. A key challenge that persists is the incorporation of spatial and temporal correlations inherent in large-scale societal event occurrences. For example, civil unrest (protests or strikes) in a city is often influenced by happenings at nearby locations and while event counts might not be comparable, there is significant temporal and spatial correlation across event occurrences.

Figure 5.1 shows a precursor event sequence discovered by our proposed model. The target event is a student protest against a French satirical magazine in the city of Bannu in Pakistan. One week before this event, there were several highly related events that occurred in other cities and may have
influenced this target event of interest. For instance, the government of Pakistan passed resolutions condemning the publication in this French magazine. Later, different groups expressed concern in small gatherings in Islamabad and Lahore followed by protests in multiple cities including Bannu.

In this chapter, we propose STAPLE, a multi-task Spatio-Temporal Precursor Learning and Event forecasting framework for multiple cities, specifically designed to discover precursors across geolocations with imbalanced class distributions and partial labels. The primary datasets of interest are open source news articles across the world encoded into events, where each event has vital information including a timestamp (in granularity of days), a geolocation (at the city level), a description (plain text), and an event type (e.g., protests). Our objective is to build forecasting models for specific cities and to identify evidential precursors from multiple cities in the past news articles.

This problem is non-trivial and poses several unique challenges: (i) *Temporal ordering constraints on events*. Events are often carefully sequenced in terms of their precursors and ignoring temporal information that is inherent in event evolution leads to unsatisfactory results. (ii) *Lack of class labels for precursor documents*. While events of interest can be manually (or automatically) detected and classified, labels for associated precursor documents (which are larger in number) are not available and are expensive to obtain. (iii) *Data scarcity and imbalanced distribution in certain geolocations*. Although a few transfer learning algorithms [66, 60, 69] support inference of the type considered here, none of them can tackle the data insufficiency problem in the presence of spatio-temporal correlations. (iv) *Inadequacy of static features*. It has been demonstrated [71] that successful event forecasting requires moving beyond static features, e.g., combining keyword frequencies with dynamic graph features. Thus the proposed forecasting models must support learning of appropriate representations inherently within its model-building.

We observe that by taking advantage of spatio-temporal event correlations within a multi-task learning framework, about 86% of the cities in our dataset have improved F1 scores compared to the best state-of-the-art algorithm. 60% of cities have more than 20% improvements in F1 score, especially for cities with less training data. We summarize the key contributions as follows:

- **Dynamic graph constraints for precursor learning and event forecasting**: Our model exploits event count correlations across multiple locations under dynamic temporal constraints for
jointly forecasting events and identifying precursors. A fusion penalty is proposed to coordinate the forecasting tasks in related cities.

- **Augmented representation learning for precursors**: By integrating document and entity embeddings within a multiple instance learning framework, it assists the prediction model to track significant entities that are evident based on their estimated probabilities.

- **Multi-task learning for precursor mining**: It alleviates the data insufficiency problem by simultaneously learning multiple related tasks and restricting all cities to share a common set of features with a consensus model.

- **Comprehensive set of experiments in real-world data**: We evaluate the proposed model on real-world datasets collected from six countries and more than one hundred cities. We conduct quantitative and qualitative analyses on the precursors inferred by the proposed model.

## 5.2 Problem Statement

Given multiple cities (or geolocations) within a country, we focus on the problem of predicting the occurrence of a future protest within a target city using captured open source news feeds. Figure 5.2 provides an overview of our proposed approach. Specifically, we hypothesize that the correlations between events occurring across “space” and “time” lead us to effectively forecast future events of
interest. Besides forecasting the protest, we also aim to identify specific news articles as precursors across different locations for manual inspection. Our proposed STAPLE model seeks to capture these correlations by jointly training the models across different cities within a nested multiple instance learning framework [43].

Formally, given a set of multiple cities $K$, each city $k$, has a set of associated news articles and event indicators ($i = 1, ..., N_k$) which are denoted by $(X^k, Y^k)$.

$$Y_i^k = \begin{cases} 
1 & \text{if an event occurs after } X_i^k \\
0 & \text{otherwise} 
\end{cases}$$  \hspace{1cm} (5.1)

The news articles published $H$ days before an indicator are given by $X_i^k$ (bag). The proposed prediction model seeks to estimate the probability of occurrence of a future event in city $k$, $P_k^i$ given $X_i^k$. Since instance-level labels are not provided, simple multi-instance learning (MIL) structures or complicated layers of MIL can both be applied in this problem. Given the reported performance [43], we use the nested multiple instance learning approach that allows for the transfer of class labels (target events) to individual news articles. It provides probabilistic estimates per document, per day, and per event for each city. The key contribution of this work is to formalize a multi-task precursor learning model to study dynamic temporal relationships of multiple geolocations. Also, we derive an optimization method to solve this problem.

5.3 Methodology

5.3.1 The Proposed Model

Given $K$ cities, let us assume $\Theta = (\theta^1, ..., \theta^K)$ be the model parameters to be learned for each of the cities, where $\theta^k \in \mathbb{R}^m$, and $m$ is the dimension of the feature space representing each document. We model the document-level probability estimates $p_{hj}$ for a news article $j$ published on day $h$ in city $k$ to associate with the event of interest using a logistic function given by $p_{hj} = \sigma(x_{hj}^T \theta^k)$ where $\sigma(a) = 1/(1 + e^{-a})$. Document embeddings ($x_{hj}$) are described in detail in Section 5.3.3. A specific document is considered to be more related to the event of interest when the probability estimate is high. Using the learned $\theta^k$ for a city, the nested multiple instance learning framework
provides probabilistic estimates by averaging at the day-level (intermediate level) and at the event-level which considers a set of consecutive days before the target event.

The STAPLE model seeks to jointly learn the different model vectors, $\Theta$, across the $K$ cities using the multi-task learning paradigm, given by:

$$
\min_{\Theta} \sum_{k \in K} \frac{N_k}{N} \mathcal{L}(\theta^k) + \lambda R(\Theta)
$$

(5.2)

where $N_k$ is the number of training examples available for city $k$ and $N$ is the total number of training examples across all the cities. $\mathcal{L}$ is the loss function and $R$ is the regularization function.

The two-level multiple instance loss function is designed to minimize the prediction error associated with forecasting events, to enforce consistency in probabilistic estimates obtained for consecutive days, and also to maximize the difference between positive and negative instances using an unsupervised hinge loss function. The regularization term explicitly captures the spatio-temporal correlations between the occurrence of events across different cities. A more specific form of the objective function corresponding to Eq. (5.2) is given below:

$$
\min_{\Theta} \sum_{k \in K} \left( \frac{N_k}{N} \mathcal{L}(\theta^k) + \frac{\lambda_1}{2} \sum_{i \in \mathcal{G}_t} \sum_{l \in \mathcal{G}_t} \alpha_{k,l}^{t_i} (\theta^k - \theta^l)^2 + \frac{\lambda_2}{2} \left\| \hat{\theta} - \theta^k \right\|_2^2 + \frac{\lambda_3}{2} \left\| \theta^k \right\|_2^2 \right)
$$

(5.3)

where $k,l$ are the indices for cities, $\theta^k$ is the model parameter for city $k$, $\hat{\theta}$ represents the global model, $t_i$ is the time index for the current event indicator, $\alpha_{k,l}^{t_i}$ is the weight between city $k$ and city $l$ at time $t_i$, $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters. Different types of penalty functions allow us to enforce different behaviors in the evolution of the event across multiple geolocations.

Within the MTL framework the regularization term is designed to enforce different penalties such that related tasks share similar features or model parameters. The STAPLE model explicitly enforces pairs of cities within a country that have seen similar events occur in the past learn similar model vectors.

$\mathcal{G}_t$ is the correlation graph for time $t_i$ that determines the tasks (cities) with similar event profiles to each other.
5.3.2 Event Correlation Graph

We demonstrate the concept of the spatio-temporal correlation graph in Figure 5.3. Each node represents a city in a country. To predict the occurrence of a protest in city A on April 1, 2015, the model analyzes the past few days of data to discover if there is an event in city A and also in other cities (i.e., city B, C and D) from Mar. 20 to Mar. 31, 2015. The weight on the edge denotes the minimum number of common events between the two cities. From Mar. 20-23, the neighbor network for city A consists of three cities B, C, D with two types of events. \( \alpha_{k,l}^{t_i} \) is the normalized weight on the edge between city \( k \) and city \( l \), given by:

\[
\alpha_{k,l}^{t_i} = \left( \sum_{c} \sum_{t=t_i-H}^{t_i} \min(E_{k,t}^i(c), E_{l,t}^i(c)) \right) + \left( \frac{1}{\text{dist}(k,l)} \right)'
\]

Here \( c \) is the event type (such as a protest), and \( E_{k,t}^i(c) \) is the number of events of type \( c \) in city \( k \) that occur within time window \( t \). We scale the value of event count and \( 1/\text{dist}() \) by feature scaling function \( (x)' = \frac{(x-x_{\text{min}})}{x_{\text{max}}-x_{\text{min}}} \) into a range of \((0.0, 1.0)\). Given the spatio-temporal correlation graph \( G \) and the edge weights, it is reasonable to assume that two cities will share several common edges, as

Figure 5.3: An example of the spatio-temporal correlation graph. Each node (A,B,C,D) represents a city. An edge between two nodes indicates that the same type of events occurred in the same time window for these two cities.
they tend to be influenced by the same set of covariates. When a city has large number of training examples, the empirical loss helps to reduce the prediction error. The spatio-temporal correlation constraint captures the task relatedness between multiple cities within a given time period. The \( \text{dist} \) function returns the distance between city \( k \) and city \( l \). We assume that two cities that are far away from each other have fewer correlations/similarities in their models.

Besides the spatio-temporal constraints we also enforce the learned individual model vectors to not deviate from the global average along with a \( l_2 \)-norm constraint on the weight vectors. The regularization parameters control the model complexity by enhancing robustness and the MTL constraints alleviate the data insufficiency problem for each individual task (if learned separately).

### 5.3.3 Learning Representations

One of the challenges in precursor mining for event forecasting is to select informative and related documents. A precursor event is not necessarily similar in semantics to the event of interest. In this work, we make use of augmented distributed representations of the documents to discover progression of precursors towards the target events. More specifically, we study the following aspects of representations for each historical news document \( x \):
Algorithm 3 STAPLE Model Learning

1: **Input**: Dataset$(X^k, Y^k), k \in [1, ... K], \tau$

2: **Output**: $\Theta = [\hat{\theta}, \theta^k]$

3: for $\tau$ iterations do

4: randomize$(X^k, Y^k)$

5: for city $k$ in $K$ do

6: for $i \in [1, 2, ..., N_k]$ do $\triangleright$ training examples for city $k$

7: graph construct $G_{t_i}$

8: calculate $\alpha_{k,l}^{t_i}$ $\triangleright$ Eq. (5.4)

9: calculate gradient $\nabla(\Theta)$

10: update $\Theta$ using $\nabla(\Theta)$ based on SGD $\triangleright$ Eq. (5.5, 5.6)

return $\Theta$

**Document Embeddings**: Recent work has shown that the semantic relationships of words can be effectively captured using the geometry of a continuous embedding space [33, 40]. For the articles in each country, we learn distributed representations (vectors) [33] for text documents and utilize these embeddings in our experiments.

**Entity Embeddings**: In many societal events, the roles of significant entities such as government officers or large organizations are substantial and sometimes even influence the progression of events. We focus on location names, person names, and organization names as the primary entities of interest. Entity and relation embeddings have been studied in structured learning and knowledge graph modeling [36]. More specifically, we use the Stanford Named Entity Recognizer (NER) [15] to extract a set of entities for English and use a series of language enrichment steps (see [46] for details) for processing. We apply the Continuous Bag-of-Word model (CBOW) [39] to each dataset to obtain word level embeddings. Finally, we aggregate the embeddings of entities into the representation for text documents as shown in Figure 5.4.
5.3.4 Optimization

The optimization problem is solved using the mini-batch gradient descent algorithm for each of the model parameters, $\theta^k$, described in Algorithm 3. We update the weight vector using an adaptive learning rate $\theta^k_{\tau+1} = \theta^k_\tau - \eta \nabla (\theta^k)$ where $\eta$ is the learning rate at the current iteration.

More specifically,

$$\theta^k \leftarrow \theta^k - \eta \left[ \frac{\partial L(\theta^k)}{\theta^k} + \lambda_1 \sum_{i \in \mathcal{G}_i} \alpha_{i,l}(\theta^k - \theta^l) - \lambda_2 (\hat{\theta} - \theta^k) + \lambda_3 \theta^k \right]$$

(5.5)

$$\hat{\theta} \leftarrow \hat{\theta} - \eta \lambda_2 (\hat{\theta} - \theta^k)$$

(5.6)

For each city, we update the global model parameters $\hat{\theta}$ and $\theta^k$ in an alternate manner. The spatio-temporal correlation graph and the weights on edges are precomputed only using the training set.

5.3.5 STAPLE Model for Precursor Mining

The precursor documents are identified based on their estimated probabilities from the learned model of each city as described in Algorithm 4 from lines 3 to 11. We also discover precursors for each city in the past time window from the neighboring cities (lines 12 to 17). For instance, in Figure 5.3, the precursors for the protest event in city A are also explored in the news articles geolocated at its neighboring cities, B to D using the learned model vectors for each of the cities. If the estimated probability is above a certain threshold, we select it into the precursor candidate pool for the target event in city A. The time complexity is dependent on the number of examples in the training dataset and the number of historical documents for each event. Based on the estimated probability of instances for events, the precursor documents and entities can be selected in linear time with respect to the number of historical documents for each event (lines 7 to 10).
Algorithm 4 STAPLE for Precursor Discovery

1: Input: Dataset $(X^k, Y^k), k \in [1, ..., K]$
2: Output: the estimated probabilities set $Q$, the discovered precursor set $O$.
3: $\Theta \leftarrow \text{call Algorithm 3 for training}$
4: for city $k$ in $K$ do
5:   for $i \in [1, 2, ..., N_k]$ do
6:     for $x_{hj} \in X_i^k$ do
7:       estimate $p_{hj}$ using $\theta^k$
8:       if $p_{hj} \geq 0.5$ then
9:         $O_i^k \leftarrow x_{hj}, Q_i^k \leftarrow p_{hj}$
10:    estimate $P_i^k$ by $\text{Avg}(p_{hj})$
11:    get “neighbor” cities for $k$ from $G_t$
12:    for city $l$ in the “neighbor” cities do
13:      for $h = [1, ..., H]$ do
14:        for $x_{hj} \in X_l^h$ do
15:          $p_{hj} = \sigma(\theta^l, x_{hj})$
16:          $O_i^k \leftarrow x_{hj}$ if $p_{hj} \geq \xi$
17: return $O, Q$

5.4 Experimental Setup

5.4.1 Datasets

We evaluated our models on event encoded data from six countries. Among them, three countries, Columbia (CO), Paraguay (PY), and Venezuela (VE) are from a labeled set called Gold Standard Report (GSR) [46] from January 2014 to April 2015. The GSR is a manually curated dataset that records civil unrest events from the ten most significant news outlets as ranked by the International Media and Newspapers in each country. An example of a recorded event is given by its city, state, country, date, and event type.
Table 5.1: Datasets used in our experiments. CO: Columbia, PY: Paraguay, VE: Venezuela, PK: Pakistan, IR: Iran, AF: Afghanistan

<table>
<thead>
<tr>
<th></th>
<th>CO</th>
<th>PY</th>
<th>VE</th>
<th>PK</th>
<th>IR</th>
<th>AF</th>
</tr>
</thead>
<tbody>
<tr>
<td># news</td>
<td>8,386</td>
<td>7,879</td>
<td>9,390</td>
<td>64,868</td>
<td>38,113</td>
<td>27,786</td>
</tr>
<tr>
<td># events</td>
<td>604</td>
<td>971</td>
<td>1,911</td>
<td>1,291</td>
<td>1,084</td>
<td>713</td>
</tr>
</tbody>
</table>

The other three datasets, Pakistan (PK), Iran (IR) and Afghanistan (AF) are from the ICEWS dataset [8] which stands for Integrated Crisis Early Warning System. It contains news articles published all over the world with the goal of evaluating national and international crisis events. Events are automatically identified and extracted by the BBN ACCENT event encoder. In our experiments, we only use data from 2015 and 2016. Each news article is labeled in one of the 20 categories. We use “protest” events as our positive examples and no protest days (or other types of events) as negative examples. We observe that the event distribution varies significantly across cities, with large cities having relatively more event occurrences compared to small cities. Statistics about these datasets are shown in Table 5.1. We use protests in the aforementioned Latin American and Middle East/Asian countries to rigorously evaluate the performance of our framework and also consider other countries (e.g., France) to provide case studies of how our framework works because these countries feature more well-known protests (e.g., climate protests in France, see Table. 6.4).

5.4.2 Experimental Protocol

We learn 300-dimensional representations for documents and 100-dimensional embeddings for words. Each document is represented by concatenating the document and entity embeddings. The entity embedding is an aggregation of each entity within the document. The GSR and ICEWS datasets record protest events on a given day at a specific location (city level). To evaluate our proposed model, for each protest event, we extract all the published reports (news articles) for up to two weeks before the occurrence of the specific event of interest. This ordered collection of per-day news documents to the protest day are considered as positive super bags. For negative examples, we identify consecutive sets of days within our studied time periods for each city when no protest event was reported. Because of the imbalanced nature of the dataset, we apply the oversampling
technique to adjust the class distribution. We sample the small (positive) class at random with replacement until it has as many samples as the majority class.

5.4.3 Comparative Methods

We compare the following methods to our proposed approach:

- **Multi-Instance SVM (MI-SVM [3])**: The MI-SVM model extends the notion of a margin from individual patterns to bags. The margin is defined between the “most positive” instance of the positive bag and the “least negative” instance of the negative bag. We collapse the news articles from the different historical days into one bag and apply this standard MIL formulation based on the SVM in scikit-learn.

- **Relaxed MIL Model (rMILavg [65])**: We estimate the probability of each instance in a bag using a logistic function. We use the average of instance-level probabilities to model the probabilities for bags.

- **Nested MIL (Nested [43])**: This approach applies a nested level of multi-instance learning for the event forecasting problem. It learns a “global” model for all cities in a country.

- **Transfer model (STAPLE-tx)**: This method is a simpler variant of the proposed model. Event forecasting for each city is considered as one task. We first apply our objective function to cities with “rich” datasets (more event examples). After learning models for these source cities, we incorporate these models into an average model with mean and standard deviation $\mu, \sigma$ and transfer them to the cities that have “sparse” datasets [51]. The target cities will then learn their models by assuming that their model parameters are drawn from a Gaussian distribution $N(\mu, \sigma)$ where $S$ denotes the set of source cities:

$$\mu = \frac{1}{K-1} \sum_{k \in S} \theta^k, \quad \sigma = \sqrt{\frac{1}{|S|-1} \sum_{k \in S} (\theta - \mu)^2}$$  \hspace{1cm} (5.7)

\hspace{1cm} ¹http://scikit-learn.org/stable/modules/svm.html
Table 5.2: F1 evaluation for the proposed methods (text embeddings vs text+entity embeddings).

<table>
<thead>
<tr>
<th>Data</th>
<th>Model</th>
<th>Text</th>
<th>Text+Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF</td>
<td>STAPLE-tx</td>
<td>0.772</td>
<td>0.783</td>
</tr>
<tr>
<td></td>
<td>STAPLE</td>
<td>0.910</td>
<td>0.910</td>
</tr>
<tr>
<td>IR</td>
<td>STAPLE-tx</td>
<td>0.880</td>
<td>0.893</td>
</tr>
<tr>
<td></td>
<td>STAPLE</td>
<td>0.895</td>
<td>0.904</td>
</tr>
<tr>
<td>PK</td>
<td>STAPLE-tx</td>
<td>0.729</td>
<td>0.725</td>
</tr>
<tr>
<td></td>
<td>STAPLE</td>
<td>0.730</td>
<td>0.713</td>
</tr>
</tbody>
</table>

5.5 Experimental Results

We perform a comprehensive empirical study to evaluate the performance of the proposed models in terms of event recall, event precision, and F1 score. We also analyze the quality of precursors with quantitative metrics and detailed case studies that highlight the strengths of the STAPLE model.

5.5.1 How well does STAPLE forecast?

Figure 5.5 shows the prediction performance of MI-SVM, rMILavg, Nested and STAPLE methods in terms of recall, precision, and F1 scores in the six datasets considered. The model hyperparameters, $\lambda_1, \lambda_2$ and $\lambda_3$ are determined using a validation set for each country. $m_0$ and $p_0$ are set to 0.5 following [43]. For other comparison models, hyperparameters are chosen following their criterion in their papers. The STAPLE method outperforms the best state-of-the-art approach (Nested) by 16% to 35% for AF, IR, and PK and 5% to 15% for CO and PY, in term of F1 scores. For CO and VE, the STAPLE-tx model outperforms other methods in terms of F1 score. The results in Table 5.2 demonstrate that the methods that take into account the entity embeddings (second column) perform better than that of using the document embeddings only (first column) for AF and IR datasets but not for PK in terms of F1 score.
Figure 5.5: Prediction performance (Recall, Precision, and F1 scores) on six datasets for the comparison methods.
5.5.2 Does the Spatio-Temporal Event Correlation Graph help?

Figure 5.6 depicts the improvements of F1 scores of the STAPLE model compared to the Nested model \((F1 \text{ Lift} = \frac{\text{STAPLE}_{F1} - \text{Nested}_{F1}}{\text{Nested}_{F1}})\) versus the number of events per city. We observe that the most significant improvements are for cities such as Qom and Kerman in Iran which are relatively small cities in their respective countries.

5.5.3 How good are the precursors?

Figure 5.7 shows the average normalized Jaccard Index of precursor documents and non-precursor documents (with respect to the target document) discovered by the STAPLE and Nested models. The normalized Jaccard index is computed as the pairwise Jaccard index, scaled relative to each event of interest. It is clear that the precursor documents have higher text-based similarity to the target events compared to the non-precursor documents as in Figure 5.7b. Figure 5.7a shows that the STAPLE model discovers more semantically related documents to the target events compared to the Nested model [43] for most of the countries.
Table 5.3: Automatically discovered precursor news stories and relevant entities for one protest event of interest. (i) Precursors for an event in one city are inferred across other cities as well. (ii) Key entities participating in the events provide explanatory power to the precursors. (iii) Probability of protest events gradually increases with the accumulation of precursor events.

<table>
<thead>
<tr>
<th>Date</th>
<th>Location</th>
<th>Precursor News Summary</th>
<th>Entity</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015-11-22</td>
<td>Toulouse</td>
<td><strong>P1.</strong> More than 10,000 people marched Saturday in the French city of Toulouse for peace and against “barbarity” a week after the devastating attacks in the capital.</td>
<td>French, Toulouse</td>
<td>0.81</td>
</tr>
<tr>
<td>2015-11-23</td>
<td>Paris</td>
<td><strong>P1.</strong> Wealthy governments and other donors need to invest more to reduce carbon emissions stemming from agriculture, said a study issued ahead of U.N. climate talks in Paris next week.</td>
<td>UN, Paris</td>
<td>0.68</td>
</tr>
<tr>
<td>2015-11-24</td>
<td>Paris</td>
<td><strong>P1.</strong> China and US have vowed to join hands with France and other parties to work toward success at the UN climate summit in Paris.</td>
<td>China, US, UN, Paris</td>
<td>0.81</td>
</tr>
<tr>
<td>2015-11-25</td>
<td>Paris</td>
<td><strong>P1.</strong> The international community must secure a binding deal against climate change at key UN talks in Paris next week, German Chancellor Angela Merkel said Wednesday.</td>
<td>German, Angela Merkel, UN, French, Aquino, Paris</td>
<td>0.83</td>
</tr>
<tr>
<td>2015-11-26</td>
<td>Paris</td>
<td><strong>P1.</strong> It comes days ahead of a major UN climate summit in Paris which aims to forge an international deal to stop global warming. <strong>P2.</strong> Pope Francis visited the world’s poorest continent to issue a clarion call for the COP21.</td>
<td>UN, Pope Francis, COP21, Paris</td>
<td>0.90</td>
</tr>
<tr>
<td>2015-11-27</td>
<td>Paris</td>
<td><strong>P1.</strong> Leaders from Russia Germany, and Europe may meet around the upcoming UN climate change conference in Paris. <strong>P2.</strong> Australia kicks off climate rallies ahead of global talks.</td>
<td>Russia, Germany, Europe, Australia, Paris</td>
<td>0.94</td>
</tr>
<tr>
<td>2015-11-28</td>
<td>Paris</td>
<td><strong>P1.</strong> Activists plan to join arms and form a “human chain” in Paris on Sunday to urge action on global warming, in a muted rally after attacks on the city.</td>
<td>UN, Paris</td>
<td>0.92</td>
</tr>
<tr>
<td>2015-11-29</td>
<td>Paris</td>
<td><strong>Protest in Paris, France:</strong> Around 4,500 activists had earlier linked hands in a peaceful protest near the site of the deadliest of the attacks, pleading for leaders to curb global warming.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.5.4 Case Studies

Table 6.4 demonstrates a case study on precursor story lines identified by the STAPLE approach. The key strength of STAPLE is its ability to leverage correlated events occurring across different cities. For each protest event, \( e \), in a city \( A \), we form a neighborhood city set based on the spatio-temporal correlation graph. Then for each city in the graph, we estimate the probabilities of news in the past week using its model for event \( e \). As long as the probability is above a certain threshold, we select the news to be a precursor candidate for this event: \( O^A_e \leftarrow x \) if \( p(x) \geq \xi \). We studied ICEWS dataset from France as a case study. In the case of a protest event in Paris, France; thousands of activists started a protest pleading leaders to stop global warming near the site of a former terror attack. Our model successfully captured the following related events leading up to this protest: A week earlier, people in Toulouse marched for the devastating attacks in Paris. Many countries announced they will be attending the upcoming Climate Change Summit (COP21) in Paris. Two days before, Australia kicked off climate rallies ahead of these global talks. One day before, activists claimed the need to march for climate change amid terror threats.

In all cases, our model captured the key change points in the precursor story lines for protest events and the key entity names (such as government officials) that played a crucial role in the development of these events.
### 5.5.5 How early can the STAPLE model forecast?

Leadtime indicates the number of days in advance that the model makes predictions and historical days denotes the number of days over which the news articles are extracted as input to the prediction algorithms. In order to study the changes in performance with and without the spatio-temporal event correlation graph constraints, we present the accuracy (ACC) score with varying lead times from 1 to 5 days for the STAPLE and the Nested models in Figure 5.8. Due to space constraints, we only depict results for PY and IR. The STAPLE model is stable and consistently performs better compared to others with varying values of leadtime.

### 5.5.6 Sensitivity Analysis

Figure 5.9 illustrates the accuracy of the proposed STAPLE model on three test dataset by varying the hyper parameters of $\lambda_1$ and $\lambda_2$ (= $\lambda_3$). $\lambda_2$ and $\lambda_3$ are chosen to be the same according to our parameter search result. The performance of different values is relatively stable and is shown for only: AF, PK, and IR countries due to the space constraints.
Figure 5.9: Sensitivity analysis on hyper-parameters $\lambda_1$ and $\lambda_2$ ($= \lambda_3$). X-axis represents the varying values of $\lambda_1, \lambda_2$ and Y-axis denotes the test accuracy.

5.6 Summary

We have presented STAPLE, a multi-task spatio-temporal correlation graph model based on a two level multi-instance learning (MIL) framework for precursor mining coupled with event forecasting. Multiple models for cities are jointly learned together and proven to be effective at both forecasting and in discovering precursors. The richness of the identified precursor events demonstrates that this will be a useful tool in dissecting event happenings. For future work, we plan to study narrative generation and entity knowledge graph extraction with multi-source datasets. We also plan to deploy our model for forecasting non-political events like disease outbreaks.
Chapter 6

When do Crowds turn Violent?

In collective behavior, some mass gatherings often underlie civil disobedience activities and as such run the risk of turning violent, causing damage to both property and people. While civil unrest is a rather common phenomenon, only a small subset of them involve crowds turning violent. How can we distinguish which events are likely to lead to violence? Using articles gathered from thousands of online news sources, we study a two-level multi-instance learning formulation, CrowdForecaster, tailored to forecast violent crowd behavior, specifically violent protests. Using data from five countries in Latin America, we demonstrate not just the predictive utility of our approach, but also its effectiveness in discovering triggering factors, especially in uncovering how and when crowd behavior begets violence.

6.1 Background

Large public crowd gatherings are common in all forms of society and some of them can lead to violence, involving damage to both property and people. Examples of such crowd gatherings include political rallies, protests, and commemorative events. When a crowd turns violent it generates economic, political, and social costs, in addition to the emotional and physical (including death) consequences for individuals directly involved in the violence. Each of the parties who support
the right to peaceful gatherings (e.g., government and police officials, community organizations) seek to develop better insights into the triggers that can foment violence in hopes of reducing the risk of violence. Efforts to decrease the probability of a violent gathering without understanding the dynamics that differentiate violent from non-violent events can lead to measures that instead increase that probability. For example, deploying a significant show of force with police and the military at the first sign of a protest can make the protesters feel intimidated and frustrated rather than protected. That frustration can, as we suggest, build into anger and increase the likelihood of violence during the next such gathering.

The outbreak of violent crowd behavior in public is usually a culmination of a stream of preceding unresolved public issues or events. As such, we hypothesize that there would be an underlying progression of events that have occurred in the past that may cause outrage and action in violence. Figure 6.1 demonstrates the average number of public gatherings that have occurred before both violent and non-violent events in five South American countries. Before the occurrence of a violent event, more protests occur on average in the prior week in comparison to a non-violent event for all the countries. Inspired by this observation, we leverage recent work in multi-instance learning [43] to develop new methods that forecast the occurrence of violent crowd behavior in advance. In particular, by integrating the correlation between the past protest events and future violent protest events, the model forecasts outbreaks of crowd violence using historical web data in both spatial and temporal aspects. Moreover, this approach identifies the precursors in the days preceding the violent protest. Our key contributions are summarized as follows:
We develop a framework based on multi-instance learning for forecasting violent protest events. The framework is built on the hypothesis that violent crowd behavior tends to have a qualitatively different set of trigger events signaling the occurrence in the future. The framework is significantly advantageous over computer vision techniques (e.g. [18]) that only detect events (not forecast them) and which require the first images of violence to be published.

In addition to forecasting violent events better, our approach leads to explainable predictions by identifying related documents in the past that can be viewed as precursors for violence. Such evidence helps policy-makers and social scientists to better understand the processes governing the formation of collective identities that turn to violence.

We conduct extensive experiments comparing our approach with existing state-of-the-art models on open datasets from five different countries. In particular, the proposed model outperforms a deployed online system (EMBERS [46, 41]), in terms of quality of forecasts and other baseline methods in terms of accuracy and AUC scores.

A civil unrest event is considered as a violent event when a news report indicates some sort of violence directly connected to the event, either directed towards people or property [43]. For example, clash with police or clash between two opposing groups that result in injuries, or significant destruction of property (e.g., burning cars, looting shops, or destruction of public property) are examples of violent events.

The specific objective of our proposed method is to forecast the occurrence of a violent event with some lead time coupled with identification of news articles reported in the past, considered as precursors leading to the specific violent event. Our analysis is conducted on a per-city level to ensure that precursors identified are localized to specific events of interest. For each city, we aggregate the set of news articles published in a day within a group (referred to as a bag). Some of these news articles are labeled as describing a protest (automatic event coders [54] are available for this task). Violence indicator discovery is formulated as a classification problem within the MIL framework wherein the aggregate collection of documents $H$ days prior to the present day are modeled as a bag of bags (one for each day), or a super bag. This super bag comprises $H$ days (bags) of news articles prior to a violent event (positive examples) or non-violent event (negative examples); with the objective of the proposed method to classify them with a lead time of $k$ days.
6.2 Methodology

Formally, we are given a set of training examples $D = \{S, Y, V\}_{i=1}^{N}$, where $S = \{X_i\}_{i=1}^{H}$ is a bag of history days and $Y \in \{0, 1\}, V \in \{0, 1\}$ when $Y = 1$. $Y$ is the label for one training example indicating if a protest event occurs on day $t+k$. $k \in [1, 2, ...]$ is the lead time of prediction that can be tuned in the experiment to evaluate how early the model can predict. $V$ is the label for a violent protest event on day $t+k$. Each day $X_i = \{x_{ij}\}_{j=1}^{n_i}$ is a set of documents. The overall architecture of the estimated probabilities are demonstrated in Fig. 6.2 and Fig. 6.3. In the nested multi-instance learning model [43], for a news article $x_{ij}$ ($j$ is the document index and $i$ is the day index), the probability of it being associated with a protest event is modeled as a logistic function: $p_{ij} = \sigma(wx_{ij})$. Here $w$ is the model parameter that is to be optimized and it has the same dimension as $x_{ij}$.
In our problem, there are two categories for the target events: violent and non-violent events. We introduce a new model parameter, $v$, for violent crowd events, with the same dimension as $x_{ij}$.

Likewise, the probability of a news article related to a violent event is defined as:

$$p^v_{ij} = \sigma(v^T x_{ij}) = \frac{1}{1 + e^{-v^T x_{ij}}} \quad (6.1)$$

Given a protest event, the historic daily probability ($P_i$) for day $t - i$ indicates how likely a protest event is going to happen on day $t + k$. The probability of this event occurring, $P(Y = 1)$, on day $t + k$ is estimated as an average vote from $H$ history days:

$$P(Y = 1) = \frac{1}{H} \sum_{i=t-H}^{t} P_i = \frac{1}{H} \sum_{i=t-H}^{t} \frac{1}{n_i} \sum_{j} p_{ij} \quad (6.2)$$

The probability of a violent protest event is modeled as a joint probability of violence and protest. Applying the Bayes rule, we get:

$$\gamma = P(V = 1, Y = 1) = P(V = 1|Y = 1) \cdot P(Y = 1)$$

$$= \left( \frac{1}{H} \sum_{i=t-H}^{t} \frac{1}{n_i} \sum_{j} p^v_{ij} \right) \cdot \left( \frac{1}{H} \sum_{i=t-H}^{t} \frac{1}{n_i} \sum_{j} p_{ij} \right) \quad (6.3)$$

Given a set of true labels $Y$ (protest) and $V$ (violent protest) for the super bags, we also know if any protest event occurs ($Y_i = 1/0$) on each history day $i$ in the same city before the target event. We can train our model by minimizing the following cost function with respect to $w$ and $v$ as:

$$\min_{w,v} \frac{\alpha}{n} \sum_{S \in S} \left[ \mathcal{L}(w,v) + \frac{1}{H} \sum_{i=t-H}^{t} Y_i \right]^2$$

$$+ \frac{1}{H} \sum_{i=t-H}^{t} g(w, X_i, X_{i-1})$$

$$+ \frac{1}{H} \sum_{i=t-H}^{t} \frac{1}{n_i} \sum_{j=1}^{n_i} h(x_{ij}, w) + R(w,v) \quad (6.4)$$

where $\mathcal{L}(*)$ represents the negative log-likelihood loss, $g(*) = (P_i - P_{i-1})^2$ is a squared loss function for two consecutive days requiring two days' probabilities to be close, and $h(*) = \max(0, m_0 - \text{sgn}(p_{ij} - p_0) w^T x_{ij})$ is the hinge loss function at the instance level where $p_0$ and $m_0$ are hyper-parameters. $R(w,v) = \frac{\beta_1}{2} ||w||^2 + \frac{\beta_2}{2} ||v||^2$ is the regularization for the model parameters. In particular, violence is an attribute or a result of protest events. Protest events can end either
Algorithm 5 CrowdForecaster algorithm

1: Input: $S = \{ (S_r, Y_r, V_r) \}$
2: Output: $\{ (w, v) \}$
3: Pre-compute $\{ \sum Y_i \mid (i = t - H, ..., t) \}$ for each event.
4: Initialize $w, v$
5: for $\tau = 1$ to $T$ do
    6:     for super bag $(S_r, Y_r, V_r)$ do
        7:     Fix $w$, update $v$ \hspace{1cm} \triangleright$Solving Eq. 6.6$
        8:     Fix $v$, update $w$ \hspace{1cm} \triangleright$Solving Eq. 6.7$
    return $\{ (w, v) \}$

peacefully or violently. The probability of a violent protest event is conditional on the probability of a protest event. Thus the negative log-likelihood for violent events is calculated as follows:

$$L(w, v) = -I(Y = 0) \log P(Y = 0) - (Y = 1) \log P(Y = 1)$$
$$- I(V = 1, Y = 1) \log P(V = 1, Y = 1)$$
$$- I(V = 0, Y = 1) \log P(V = 0, Y = 1)$$ (6.5)

From our observations, violent protest events usually follow a sequence of protest events. Thus, the probabilities of violent protest events are assumed to be highly related with the number of protest events that occur before these violent events ($(\gamma - \frac{1}{H} \sum_{i=t-H}^{t} Y_i)^2$). For instance, if there are seven protest events in the same city in one week, it is highly probable that some protest will turn to violence because the growing anger and dissatisfaction tend to make people resort to violence. We develop an alternating minimization algorithm that can be applied to achieve a solution of Eq. 6.4.

**Update $v$ when $w$ is fixed.** All $v$’s can be solved by a stochastic gradient descent algorithm as:

$$g'(v) = \alpha \left[ - \frac{V - \gamma}{\gamma(1-\gamma)} \frac{\partial \gamma}{\partial v} + 2 \left( \gamma - \frac{1}{H} \sum_{i=t-H}^{t} Y_i \right) \frac{\partial \gamma}{\partial v} \right] + \beta_2 v$$ (6.6)
Update $w$ when $v$ is fixed. Likewise, when $v$ is fixed, all $w$'s derivatives have a formulation as follows:

$$
g'(w) = \alpha \left[ -\frac{Y - P}{P(1 - P)} \frac{\partial P}{\partial w} - \frac{V - \gamma}{\gamma(1 - \gamma)} \frac{\partial \gamma}{\partial w} 
+ 2(\gamma - 1) \frac{1}{H} \sum_{i=t-H}^{t} Y_i \frac{\partial \gamma}{\partial w} 
+ \frac{1}{H} \sum_{i=t-H}^{t} \left( 2(P_i - P_{i-1}) \left( \frac{\partial P_i}{w} - \frac{\partial P_{i-1}}{w} \right) 
- \frac{1}{n} \sum_{j=1}^{n_i} \text{sgn}(p_{ij} - p_0) x_{ij} I_{ij} \right) \right] + \beta_1 w
$$

where $I_{ij}$ is the indicator function when $\text{sgn}(p_{ij} - p_0) w^T x_{ij} \leq m_0$, it returns 1. We alternatively updating $w$ and $v$ via gradient descent toward convergence. A complete algorithm is described in Algorithm 5. After learning the model parameters $(w, v)$ from training examples, test examples are evaluated in terms of the metrics described later in Section 6.3. One issue in the problem is the imbalanced class distribution due to fewer examples of violent protest in the real world datasets. We apply one of the traditional techniques, *OverSampling*, to adjust the class distribution. It samples the smaller class at random with replacement until it has as many samples as the majority class.

In order to discover the historical related documents for violent protest events, we apply the algorithm described in [43]. It selects instances $(x_{ij})$ as the precursor documents for each violent event based on their estimated probabilities $(p_{ij}^v)$ if the estimated $p_{ij}^v$ is beyond a threshold.

### 6.3 Experimental Evaluation

#### 6.3.1 Experimental Design.

**Datasets**

We collected news articles from top news sources in different countries including *Argentina, Brazil, Colombia, Paraguay, Venezuela* from January 2014 to April 2015. Each dataset contains about 9400 to 11,000 news and 600 to 2000 events. Among these events, 6% to 30% are violent events.
Ground Truth

The civil unrest forecasting results were validated against a labeled set called Gold Standard Report (GSR) that was exclusively provided by MITRE (see [46] for more details). The GSR is a manually curated dataset that records the occurrence of a civil unrest events reports from the ten most significant news outlets as ranked by International Media and Newspapers in each of the countries studied here. An example of a ground truth GSR recorded event is given by a tuple: (Location= “Argentina, Fortaleza, Ceara”, DATE = “2014-01-20”, Protest = “True”, Violence=“True”). Here we use the “Violence” attribute as our violence label \( V \) and “Protest” as our event label \( Y \). These GSR reports are the target events that are used for validation of our algorithm.

Experimental Setup

To evaluate the MIL-based violent event forecasting, for each violent protest event in a city, we download all the published news articles in that city for up to 14 days before the occurrence of the specific event. This ordered collection of per-day news documents up to the violent protest day are considered as positive super bags. We generate negative samples in two ways: (i) For each location (city) we identify a period of three consecutive days where no identified event of interest (as reported by GSR) occurs. The ordered collection of per-day news documents not leading to a protest event are labeled as negative super bags. (ii) Similarly for each location (city) the ordered collection of per-day news documents leading to a protest event but of non-violent nature are labeled as negative super bags as well. We split our datasets into training and testing (held-out) partitions and perform 3-fold cross-validation on the training set to tune the parameters of the proposed models. We represent documents by word and document embedding generated by paragraph vector models [33]. For each document, we learn its representation with dimension of 300 for training. According to the nMIL model, doc2vec has better performance compared to TFIDF representations.

Comparative Methods

We compare the following methods and systems in our experiments:
- **MI-SVM** [3] (MI-SVM): The MI-SVM model extends the notion of a margin from individual patterns to bags. In our case, we collapse the news articles from the \( r \) historical days into one bag and each bag has two labels indicating the occurrence of a protest event and a violent protest event on the \( t+k \)-th day. The MISVM iterates each example in each bag to determine the most positive instance and least negative instance.

- **Relaxed-MIL** [65] (Relaxed): This proposed model uses a Noisy-OR function \( (P_i = 1 - \prod_{j=1}^{m_i}(1 - p_{ij})) \) to estimate the probability of a bag being positive. In our dataset, each bag (day) has more than 10 instances (news articles). Noisy-OR function tends to generate a positive probability when the number of instances in a bag is large. Thus, we estimate the probability of a bag being positive by applying an average function of each instance in the bag.

- **Nested MIL** (nMIL): This approach is proposed by Ning et. al. [43]. Instead of general protest event prediction, we use violent protest events as the positive examples in this method.

- **Nested MIL-MC** (Multi-Class): This is the multi-class classifier of nMIL model. In this experiment, we divide the examples into three classes: violent protest, non-violent protest, and no protest.

- **EMBERS**[46] (EMBERS): This is an automated, 24x7 continuous system for forecasting civil unrest across 10 countries in Latin America using open source indicators such as tweets, news sources, blogs, economic indicators, and other data sources.

### Parameter Settings

For CrowdForecaster, the parameters were set as follows: The supervised hyper parameter \( \alpha \) is set to 0.6. The hyper parameter \( \beta_1 \) and \( \beta_2 \) were set to 0.5 according to the nMIL paper [43]. The learning rate is adaptively set to \( \frac{1}{(t+1)\cdot \lambda} \) where \( \lambda \) is 0.05 according to the rMIL paper [65]. \( m_0 \) and \( p_0 \) are set to be 0.5. Lead time is used to evaluate how early the model forecasts. For instance, if we use data from day 1 to day 5 and forecast if there is a violent event on day 6, the lead time is 1 day. In the experiment, we set lead time as 1 day by default for offline evaluation and 4 days for online evaluation. We vary the lead time from 1 day to 4 days to compare the early forecasting power of the different models.
Table 6.1: Violent event forecasting performance comparison based on Accuracy (Acc) and AUC score w.r.t to state-of-the-art methods.

The proposed CrowdForecaster method outperforms state-of-the-art methods across the five countries with 2 weeks historical data.

<table>
<thead>
<tr>
<th>Method</th>
<th>Argentina ACC</th>
<th>Argentina AUC</th>
<th>Brazil ACC</th>
<th>Brazil AUC</th>
<th>Colombia ACC</th>
<th>Colombia AUC</th>
<th>Paraguay ACC</th>
<th>Paraguay AUC</th>
<th>Venezuela ACC</th>
<th>Venezuela AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI-SVM</td>
<td>0.307</td>
<td>0.572</td>
<td>0.221</td>
<td>0.515</td>
<td>0.317</td>
<td>0.564</td>
<td>0.281</td>
<td>0.509</td>
<td>0.306</td>
<td>0.518</td>
</tr>
<tr>
<td>Relaxed</td>
<td>0.631</td>
<td>0.552</td>
<td>0.462</td>
<td>0.518</td>
<td>0.631</td>
<td>0.552</td>
<td>0.569</td>
<td>0.597</td>
<td>0.438</td>
<td>0.489</td>
</tr>
<tr>
<td>nMIL</td>
<td>0.669</td>
<td>0.568</td>
<td>0.552</td>
<td>0.522</td>
<td>0.680</td>
<td>0.645</td>
<td>0.650</td>
<td>0.589</td>
<td>0.560</td>
<td>0.570</td>
</tr>
<tr>
<td>Multi-Class</td>
<td>0.544</td>
<td>0.551</td>
<td>0.295</td>
<td>0.523</td>
<td>0.626</td>
<td>0.664</td>
<td>0.531</td>
<td>0.587</td>
<td>0.234</td>
<td>0.457</td>
</tr>
<tr>
<td>CrowdForecaster</td>
<td><strong>0.804</strong></td>
<td><strong>0.712</strong></td>
<td><strong>0.762</strong></td>
<td><strong>0.540</strong></td>
<td><strong>0.791</strong></td>
<td><strong>0.681</strong></td>
<td><strong>0.892</strong></td>
<td><strong>0.661</strong></td>
<td><strong>0.584</strong></td>
<td><strong>0.594</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Argentina (↑ %)</th>
<th>Brazil (↑ %)</th>
<th>Colombia (↑ %)</th>
<th>Paraguay (↑ %)</th>
<th>Venezuela (↑ %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI-SVM(baseline)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Relaxed(↑ %)</td>
<td>106</td>
<td>-3.6</td>
<td>108.9</td>
<td>0.6</td>
<td>99.1</td>
</tr>
<tr>
<td>nMIL(↑ %)</td>
<td>118.3</td>
<td>-0.7</td>
<td>149.8</td>
<td>1.4</td>
<td>114.5</td>
</tr>
<tr>
<td>Multi-Class(↑ %)</td>
<td>77.4</td>
<td>-3.7</td>
<td>33.4</td>
<td>1.5</td>
<td>97.6</td>
</tr>
<tr>
<td>CrowdForecaster(↑ %)</td>
<td><strong>162.5</strong></td>
<td><strong>24.5</strong></td>
<td><strong>244.8</strong></td>
<td><strong>4.7</strong></td>
<td><strong>149.6</strong></td>
</tr>
</tbody>
</table>
Table 6.2: AUC scores of the proposed method and the best baseline with lead time from 1 to 4 for violent events.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Leadtime(→)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>nMIL</td>
<td>0.568</td>
<td>0.608</td>
<td>0.610</td>
<td>0.656</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CrowdForecaster</td>
<td><strong>0.712</strong></td>
<td><strong>0.674</strong></td>
<td><strong>0.646</strong></td>
<td><strong>0.689</strong></td>
<td></td>
</tr>
<tr>
<td>Brazil</td>
<td>nMIL</td>
<td>0.522</td>
<td>0.519</td>
<td>0.507</td>
<td>0.573</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CrowdForecaster</td>
<td><strong>0.540</strong></td>
<td><strong>0.584</strong></td>
<td><strong>0.540</strong></td>
<td><strong>0.613</strong></td>
<td></td>
</tr>
<tr>
<td>Colombia</td>
<td>nMIL</td>
<td>0.645</td>
<td>0.549</td>
<td>0.693</td>
<td>0.627</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CrowdForecaster</td>
<td><strong>0.681</strong></td>
<td><strong>0.619</strong></td>
<td><strong>0.735</strong></td>
<td><strong>0.614</strong></td>
<td></td>
</tr>
<tr>
<td>Paraguay</td>
<td>nMIL</td>
<td>0.589</td>
<td>0.670</td>
<td>0.596</td>
<td>0.593</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CrowdForecaster</td>
<td><strong>0.661</strong></td>
<td><strong>0.758</strong></td>
<td><strong>0.635</strong></td>
<td><strong>0.692</strong></td>
<td></td>
</tr>
<tr>
<td>Venezuela</td>
<td>nMIL</td>
<td>0.570</td>
<td>0.597</td>
<td>0.609</td>
<td>0.563</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CrowdForecaster</td>
<td><strong>0.594</strong></td>
<td><strong>0.628</strong></td>
<td><strong>0.642</strong></td>
<td><strong>0.588</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Performance Metrics**

The forecasting alerts generated by the model and the real events are structured records as follows:

| Date: 2014-07-01, Type: Violent protest. | Location: Brazil, Sao Paulo, Sao Paulo. |

The quality score for a forecast involved evaluations based on time and location given by:

\[ QS = 2 \times (DS + LS) \]

where \( DS \), \( LS \) denote the date score and location score respectively.

\[ DS = 1 - \min(|d_e - d_p|, 7)/7 \]

where \( d_e \) is the event date and \( d_p \) is the predicted date for the event. If the predicted date of the event is the same as the actual date of the event, then \( DS \) is 1. Location score has many ways of definition. In our problem, location is in terms of a triples of (country, state, city). Comparing a
true event with a predicted event, we obtain a score at these three levels:

\[ LS = \frac{1}{3}l_1 + \frac{1}{3}l_1l_2 + \frac{1}{3}l_1l_2l_3 \]

where \( l_1 \) is the country-level score, \( l_2 \) is the state level score, and \( l_3 \) is the city level score. We selected a set of cities based on their scales. Then we built training and testing examples for these cities and the location score is only calculated for the selected cities.

Other typical evaluation metrics for classification include: accuracy (ACC) and area under curve (AUC) score. True positive examples are the true violent events and the model predicts correctly. True negative examples are the true non-violent events or no-event and model predicts correctly.

### 6.3.2 Experimental Results.

We introduce the results of forecasting violent crowd behavior in several parts. First, we show the offline performance comparison with models MI-SVM, Relaxed, nMIL, Multi-Class based on accuracy (ACC) and AUC score for five countries in South America. Second, we present the online evaluation of quality scores (QS, LS, DS) for the proposed model and EMBERS that delivers warnings.

### Table 6.3: Quality scores for the proposed method and the delivery from online system, EMBERS [46]. DS and LS are over 1; QS is over 4.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Methods</th>
<th>DS</th>
<th>LS</th>
<th>QS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>EMBERS</td>
<td>0.83</td>
<td>0.69</td>
<td>3.04</td>
</tr>
<tr>
<td></td>
<td>CrowdForecaster</td>
<td>0.99</td>
<td>0.93</td>
<td>3.84</td>
</tr>
<tr>
<td>Brazil</td>
<td>EMBERS</td>
<td>0.85</td>
<td>0.81</td>
<td>3.32</td>
</tr>
<tr>
<td></td>
<td>CrowdForecaster</td>
<td>0.99</td>
<td>0.99</td>
<td>3.96</td>
</tr>
<tr>
<td>Colombia</td>
<td>EMBERS</td>
<td>0.82</td>
<td>0.75</td>
<td>3.14</td>
</tr>
<tr>
<td></td>
<td>CrowdForecaster</td>
<td>0.94</td>
<td>0.99</td>
<td>3.86</td>
</tr>
<tr>
<td>Paraguay</td>
<td>EMBERS</td>
<td>0.89</td>
<td>0.76</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>CrowdForecaster</td>
<td>0.95</td>
<td>1</td>
<td>3.9</td>
</tr>
<tr>
<td>Venezuela</td>
<td>EMBERS</td>
<td>0.82</td>
<td>0.8</td>
<td>3.24</td>
</tr>
<tr>
<td></td>
<td>CrowdForecaster</td>
<td>0.93</td>
<td>0.99</td>
<td>3.84</td>
</tr>
</tbody>
</table>
24/7 for these countries. Next, we study a few cases of precursor discovery for violent and non-violent events. We also calculate and analyze the computation time of the proposed model with other models. Finally, we present the sensitivity of model parameters on the performance of the proposed method.

**The overall accuracy for violence prediction**

Table 6.1 lists the comparative performance in terms of Accuracy (ACC) and AUC scores for five countries in South America. In the bottom part of the table, the lift/drop percentage of the proposed model is presented comparing to the MI-SVM as a baseline. The accuracy computes the fraction of correct predictions for both positive and negative classes. The AUC is a common evaluation metric for binary classification problems with imbalance. It will be close to 1 when the true positive rate increases quickly. The proposed model outperforms other state-of-the-art methods for all datasets with lead time equal to one day. With a portion of training samples changing from 10% to 100%, Figure 6.4 shows the AUC scores for the CrowdForecaster model and other state-of-the-art models for Argentina dataset. Given the space limitation, we only show this result on Argentina. In general, the AUC scores increase when the number of training samples is increased. With the full set of training samples, the proposed CrowdForecaster model outperforms MI-SVM, Relaxed and nMIL methods by 26%, 25% and 22%, respectively.

**Leadtime Evaluation**

Table 6.2 shows the AUC performance of the proposed method CrowdForecaster in comparison to the best baseline method, nMIL, with lead time varying from 1 to 4. For each value of the lead time we train a model where $X_r$ is a super bag containing $t$ historical days and $Y_r$ indicates if a violent protest event happened on day $t+k$. Notice that lead time $k$ indicates the model predicts $k$ days in advance. For Brazil, Colombia, Paraguay and Venezuela, it is noted that a shorter lead time ($k=1$) does not necessarily imply a better predictive performance compared to a longer lead time ($k=3, 4$).
Comparison to the online system in production

Table 6.3 presents the performance with respect to quality scores for EMBERS delivered system and the proposed methods. The proposed model, CrowdForecaster, outperforms the online system in event data and event location scores for all the five datasets with an average performance improvement of 21%.

Trigger Analysis from Results

Table 6.4 presents case studies on two violent protests and one non-violent protest. The detected precursor events are reported before these protest events. We select each precursor event by setting a threshold for its probability $p_{ij}^v > 0.5$. The top words are selected from the precursor document of that day based on their frequency. We can make three observations about these results.
First, the occurrence of keywords such as “police”, “teargas” in the precursors for violent protests suggest that greater forms of control and authority often beget violence. This reflects ongoing understanding that violence often has at its roots distrust between authorities and citizens. Second, we observe that before a violent protest event there have been other protests and strike events (even if peaceful) among the precursors. Words such as “protest”, “block”, “march”, “strike” tend to appear more frequently in news articles preceding a violent protest than a non-violent protest. These findings suggest that a rapidly rising sequence of peaceful events can produce an emotional state among protesters that increases the likelihood of violence in the next protest. We suggest that emotional state is frustration. Given the myriad of challenges confronting civil society in Latin America today, fortunately frustration is not sufficient to produce violence. But the pattern of an increase in a set of words and events is intriguing. Third, analysis of precursor events indicates that they are occurring across a few cities or states, not just in the locale that will subsequently experience violence. One might have thought that protesters would be most affected by what happens locally but this data suggests that those protesters prone to violence reflect upon their national and not just local experiences when formulating their grievances and developing their frustrations. This correlation needs to be explored carefully because it both limits the responsibility of local authorities for potential violence and suggests that locally focused tactics to lower the risk of violence will be of little value.

**Trigger Analysis from Twitter**

In order to discover related topics and emotions from social media, we conduct a post-prediction analysis from Twitter. Firstly, we collect tweets by a seed set of keywords including “march”, “protest”, and “street”. Then we expand the keyword set by using dynamic query expansion to get more tweets. We filter the tweets by their locations and timestamps. For each country, we first select the baseline keywords as follows: for each city where violent events occurred, we calculate the daily word frequency from tweets and get averaged values as shown in Table 6.5 and Table 6.6. For each violent event (VE), we calculate the daily word frequency in each of the preceding 3 days at the event city. Then we get averaged word frequencies for all violent events in each country.

As seen in Table 6.5 and 6.6, we observe that the word distributions before violent events are different from baseline, and that topics vary across countries. For instance, in Argentina, three
days before violent events, people talk more about “jobs”, “food”, “security”, and “death”. While in Venezuela, topics such as “students”, “crisis”, “gasoline”, and “complaint” draw more attention in Twitter.

Computational Complexity

Figure 6.5 presents the computation time of the proposed method and state-of-the-art methods on a Dell server with Intel Xeon CPU, 80-core, 504 GB memory based Ubuntu 12.04.5 operating system. The MI-SVM algorithm is computationally more expensive when the number of training examples is large. Other probability based MIL methods are relatively stable when the training set is varied from 10% to 100%.

Sensitivity Analysis for Hyperparameters

We also assess the proposed model with different values of parameters. Figure 6.6 shows the relatively stable forecasting performance for Argentina dataset varying $\alpha$ and $\beta$. 
6.4 Summary

We have introduced a framework based on multi-instance learning for forecasting violent crowd behavior, and identifying precursor events that lead to violence. We empirically evaluated the strengths of our developed method on open source news datasets from five Latin American countries and we conducted a semantic analysis using social media for violent events. Through extensive evaluation and analysis, we illustrate the strong forecasting performance of the proposed methods for violence prediction. We also show qualitatively, via several case studies, the characteristics of identified precursors for both violent and non-violent protest events. In the future, we plan to study the patterns of change in protest events that turn to violence and other societal factors that contribute to the evolution of violent protest events.
Table 6.4: Precursor events and word distributions in past seven days for violent protest events and one non-violent protest event. Related keywords are manually highlighted (from Google Translate).

<table>
<thead>
<tr>
<th>Day</th>
<th>Precursor Events</th>
<th>Top Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day T-8</td>
<td>P1. The rain of Sao Paulo didn’t prevent women march on street for equal rights</td>
<td>street</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day T-5</td>
<td>P1. Protest in front of the central railway station of Brazil.</td>
<td>protest</td>
</tr>
<tr>
<td></td>
<td>P2. About 25 students and their fathers participate a protest</td>
<td>students</td>
</tr>
<tr>
<td>Day T-3</td>
<td>P1. Thousands of protesters gather together in the Cinelandia, downtown Rio.</td>
<td>protest</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day T-2</td>
<td>P1. The demonstration against the Dilma government and corruption in Belo Horizonte.</td>
<td>government</td>
</tr>
<tr>
<td></td>
<td>P2. About 3000 people, according to estimate of the military police, participate in the protest</td>
<td>police</td>
</tr>
</tbody>
</table>

2015-03-17 Violent Protest: A group of protesters close the runway in the marginal Tiete, burn tires and garbage bags.

<table>
<thead>
<tr>
<th>Day</th>
<th>Precursor Events</th>
<th>Top Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day T-8</td>
<td>P1. A young guy died from an attempt of kidnapping.</td>
<td>criminal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>justice</td>
</tr>
<tr>
<td>Day T-7</td>
<td>P1. Senate start debate on the reform of the political penal code and the criminal justice commission matter</td>
<td>economic</td>
</tr>
<tr>
<td></td>
<td>P2. Deputy of the New Alliance Party arrive to the session of the plenary of the local congress,</td>
<td>maintain</td>
</tr>
<tr>
<td>Day T-5</td>
<td>P1. The workers against Congress gather in May Square.</td>
<td>human</td>
</tr>
<tr>
<td>Day T-2</td>
<td>P1. The judge rejected the proposition on one of the accused criminals who took the property of Lugano</td>
<td>financial</td>
</tr>
<tr>
<td></td>
<td></td>
<td>property</td>
</tr>
<tr>
<td>Day T-1</td>
<td>P1. Men in police disguise assaulted the house of a doctor.</td>
<td>medical</td>
</tr>
<tr>
<td></td>
<td>P2. After two hours of chaos, workers lead a protest</td>
<td>police</td>
</tr>
</tbody>
</table>

2014-05-07 Non-Violent Protest: About 500 people gathered last night at the Plaza Santo Martin to demand the authority to enforce the security of the citizens.
Table 6.5: Top Twitter words in countries: Argentina, Brazil, and Columbia. Color-highlighted words are those that appear more before violent events than normal days. Blue: government related topics; Green: actions and emotions; Red: other topics.

<table>
<thead>
<tr>
<th>Argentina</th>
<th>Brazil</th>
<th>Colombia</th>
</tr>
</thead>
<tbody>
<tr>
<td>power</td>
<td>hunger</td>
<td>deputies</td>
</tr>
<tr>
<td>before VE</td>
<td>before VE</td>
<td>before VE</td>
</tr>
<tr>
<td>119</td>
<td>143</td>
<td>23</td>
</tr>
<tr>
<td>hate</td>
<td>fear</td>
<td>people</td>
</tr>
<tr>
<td>38</td>
<td>49</td>
<td>16</td>
</tr>
<tr>
<td>fear</td>
<td>money</td>
<td>family</td>
</tr>
<tr>
<td>106</td>
<td>141</td>
<td>22</td>
</tr>
<tr>
<td>power</td>
<td>time</td>
<td>time</td>
</tr>
<tr>
<td>33</td>
<td>43</td>
<td>11</td>
</tr>
<tr>
<td>job</td>
<td>power</td>
<td>job</td>
</tr>
<tr>
<td>60</td>
<td>129</td>
<td>10</td>
</tr>
<tr>
<td>fear</td>
<td>followers</td>
<td>followers</td>
</tr>
<tr>
<td>30</td>
<td>41</td>
<td>9</td>
</tr>
<tr>
<td>food</td>
<td>job</td>
<td>unrest</td>
</tr>
<tr>
<td>50</td>
<td>117</td>
<td>9</td>
</tr>
<tr>
<td>followers</td>
<td>money</td>
<td>family</td>
</tr>
<tr>
<td>21</td>
<td>39</td>
<td>9</td>
</tr>
<tr>
<td>kill</td>
<td>change</td>
<td>mayor</td>
</tr>
<tr>
<td>48</td>
<td>104</td>
<td>7</td>
</tr>
<tr>
<td>family</td>
<td>hunger</td>
<td>fear</td>
</tr>
<tr>
<td>21</td>
<td>37</td>
<td>8</td>
</tr>
<tr>
<td>dies</td>
<td>kill</td>
<td>force</td>
</tr>
<tr>
<td>39</td>
<td>79</td>
<td>6</td>
</tr>
<tr>
<td>people</td>
<td>power</td>
<td>power</td>
</tr>
<tr>
<td>20</td>
<td>36</td>
<td>8</td>
</tr>
<tr>
<td>gov</td>
<td>food</td>
<td>president</td>
</tr>
<tr>
<td>39</td>
<td>75</td>
<td>6</td>
</tr>
<tr>
<td>hungry</td>
<td>job</td>
<td>problem</td>
</tr>
<tr>
<td>19</td>
<td>32</td>
<td>7</td>
</tr>
<tr>
<td>security</td>
<td>rage</td>
<td>gov</td>
</tr>
<tr>
<td>37</td>
<td>74</td>
<td>6</td>
</tr>
<tr>
<td>death</td>
<td>change</td>
<td>water</td>
</tr>
<tr>
<td>19</td>
<td>31</td>
<td>6</td>
</tr>
<tr>
<td>summoned</td>
<td>death</td>
<td>terrorists</td>
</tr>
<tr>
<td>36</td>
<td>59</td>
<td>4</td>
</tr>
<tr>
<td>job</td>
<td>play</td>
<td>job</td>
</tr>
<tr>
<td>18</td>
<td>27</td>
<td>6</td>
</tr>
<tr>
<td>mayor</td>
<td>president</td>
<td>violence</td>
</tr>
<tr>
<td>33</td>
<td>54</td>
<td>4</td>
</tr>
<tr>
<td>water</td>
<td>send</td>
<td>matches</td>
</tr>
<tr>
<td>17</td>
<td>24</td>
<td>5</td>
</tr>
<tr>
<td>rights</td>
<td>campaign</td>
<td>arrest</td>
</tr>
<tr>
<td>19</td>
<td>46</td>
<td>4</td>
</tr>
<tr>
<td>change</td>
<td>remember</td>
<td>mayor</td>
</tr>
<tr>
<td>17</td>
<td>23</td>
<td>5</td>
</tr>
<tr>
<td>justice</td>
<td>poor</td>
<td>fight</td>
</tr>
<tr>
<td>18</td>
<td>43</td>
<td>3</td>
</tr>
<tr>
<td>matches</td>
<td>help</td>
<td>support</td>
</tr>
<tr>
<td>17</td>
<td>23</td>
<td>5</td>
</tr>
<tr>
<td>unbearable</td>
<td>complain</td>
<td>conflict</td>
</tr>
<tr>
<td>17</td>
<td>42</td>
<td>3</td>
</tr>
<tr>
<td>kill</td>
<td>food</td>
<td>hungry</td>
</tr>
<tr>
<td>15</td>
<td>23</td>
<td>5</td>
</tr>
</tbody>
</table>
### Table 6.6: Top Twitter words in countries: Paraguay and Venezuela.

<table>
<thead>
<tr>
<th>Paraguay</th>
<th>Venezuela</th>
</tr>
</thead>
<tbody>
<tr>
<td>before VE</td>
<td>baseline</td>
</tr>
<tr>
<td>faillures</td>
<td>6 followers</td>
</tr>
<tr>
<td>fear</td>
<td>3 movement</td>
</tr>
<tr>
<td>violence</td>
<td>3 accident</td>
</tr>
<tr>
<td>assassin</td>
<td>3 merchants</td>
</tr>
<tr>
<td>attack</td>
<td>2 results</td>
</tr>
<tr>
<td>hate</td>
<td>2 dictator</td>
</tr>
<tr>
<td>president</td>
<td>2 guests</td>
</tr>
<tr>
<td>power</td>
<td>2 fire</td>
</tr>
<tr>
<td>armed</td>
<td>2 aggression</td>
</tr>
<tr>
<td>perceive</td>
<td>2 poverty</td>
</tr>
<tr>
<td>minister</td>
<td>2 crime</td>
</tr>
<tr>
<td>security</td>
<td>2 community</td>
</tr>
<tr>
<td>taxi</td>
<td>1 warning</td>
</tr>
</tbody>
</table>
Chapter 7

Conclusion and Future Work

A key goal of this dissertation is to develop innovative solutions for real-world event modeling problems. In our work, we have successfully applied machine learning techniques such as multi-instance learning, multi-task learning, transfer learning, and topic modeling in event modeling to our application context.

To address the problem of information interaction between two media sources, a novel unsupervised framework has been presented for uncovering information reciprocity between news media and social media. Extensive empirical studies on story chains of Brazilian mass protests were conducted with corresponding Twitter data collected for each news article in the story chains. The results reveal our discovery in a few dimensions such as interaction patterns for different event categories, topic variability in interaction patterns, the main influencer in story development.

To discover precursor storylines in event development, a new nested structure of multi-instance learning was developed and evaluated to solve the simultaneous problems of event forecasting in city level and intelligent precursor selection for target events. Experimental evaluation was performed on news documents and manually labeled events from top news agencies in three countries in South America. The proposed approach is compared with five existing state-of-the-art methods on the predictive performance. Case studies including protests in Argentina and Mexico have demonstrated the effectiveness of precursor storylines discovered in the model.
To conquer data sparsity problems in multi-scale locations, a personalized multi-task learning method is proposed to forecast spatio-temporal events and provide multi-location precursors. It tackles the problem of sparse data distribution and reinforces co-occurring location-specific precursors with augmented representations. Through studies on three countries in South America from Gold Standard Report dataset and three countries in Asia from ICEWS dataset, the model is proven with promising performance on predictive values and precursor mining from multiple locations.

To address the problem of violent event detection, a tailored solution with multi-instance learning is introduced to trace when the events are becoming violent and the corresponding factors as evidence. Experiments are conducted on five countries in South America over one year. It is shown that the proposed model outperforms other existing models in prediction scores.

To investigate large-scale event distribution over decades and multiple regions, a systematic framework and case study is conducted on a comprehensive event dataset. This work demonstrates the strengths of the proposed model on country level precursor analysis with dynamic topic modeling. Future work will be extended to collective behavior analysis using data science and machine learning techniques.

7.1 Contributions

The major research tasks performed and completed in this thesis are listed below.

7.1.1 Information Reciprocity between News Media and Social Media

- Development of an unsupervised framework for story chaining algorithm with two media sources. This framework is based on weighted scores of similarities across news articles for textual features, spatial features, and actors.

- Design of a mechanism to classify the interaction patterns between News articles and Twitter. This encoding method is applied to all articles in a story chain resulting in a string of interaction states.
• **Clustering on interaction patterns without semantic information.** Dynamic time warping is adopted to identify clusters of the quantitative encoding for interaction patterns.

• **Topical analysis based on interactions patterns.** The primary source of information is identified based on the interaction states and topic modeling is adapted to discern content difference for different interaction patterns.

### 7.1.2 Multi-Instance Learning for Modeling Precursors and Forecasting Events

• **A novel nested multi-instance learning framework for event forecasting and precursor mining.** The event forecasting and precursor mining for multiple cities in a country are formulated as a multi-instance learning problem with a nested structure. By estimating a prediction score for each instance in the history data, significant precursors for different events are automatically detected.

• **Harnessing temporal constraints.** Different penalty functions and regularizations are explored while employing the temporal information in our dataset under assumption that most events of interest are follow-up reports of other events that happened before, and most planned events are developing over time.

• **Modeling of various event categories in multiple geo-locations.** For general purpose multi-class classification, the nested MIL formulation is extended to determine necessary attributes of events in terms of their underlying population.

• **Comprehensive experiments in real-word datasets.** The evaluation is conducted using news data collected from July 2012 to December 2014 in three countries of Latin America.

### 7.1.3 Multi-Task Learning for Modeling Spatio-Temporal Precursors and Events

• **A new multi-task learning framework to discover spatio-temporal precursors.** It alleviates the data insufficiency problem by simultaneously learning multiple related tasks and restricting all cities to share a common set of features with a consensus model.
• **Dynamic spatio-temporal graph constraints.** It exploits event count correlations across multiple locations under dynamic temporal constraints for jointly forecasting events and identifying precursors. A fusion penalty is proposed to coordinate the forecasting tasks in related cities.

• **Augmented representation learning for precursors.** By integrating document and entity embeddings within a multiple instance learning framework, it assists the prediction model to track significant entities that are evident based on their estimated probabilities.

• **Comprehensive set of experiments in real-world data.** The proposed models are evaluated for data involving six countries and more than one hundred cities.

### 7.1.4 Modeling Violent Event Progression

• **Design of a multi-instance learning framework tailored for violent event detection.** The framework presented here is built on the hypothesis that violent crowd behavior tend to have a qualitatively different set of trigger events signaling the occurrence in the future.

• **Extensive experiments in real-world datasets compared to existing state-of-the-art methods.** We conduct extensive experiments comparing our approach with existing state-of-the-art models on open datasets from five different countries.

• **Comprehensive analysis of triggers from both news media and Twitter.**

### 7.2 Future Directions

#### 7.2.1 Big Data for Computational Social Science

Online social networks such as Twitter and Facebook involve millions, even billions, of users, texts, and images. A major short-term objective is to develop scalable methods based on distributed platforms such as Spark and Hadoop to take advantage of multiple and heterogeneous data sources for detecting events and precursors. In the long term, we plan to (1) study and understand the dynamics of multi-modal data sources, (2) contribute to the development of multi-modal frameworks
in social science, and (3) build efficient and effective solutions that aim to improve the integration of data at scale.

7.2.2 Interpretability of Models in the Era of Deep Learning

This is a trending area in the context of providing user-friendly services. To provide explanations in online services, we are interested in designing interpretable models that excel in terms of performance, scalability, ease-of-use/deployment, and efficiency. Such studies will impact the next-generation of artificial intelligence and provide best practice guidance for algorithm deployment in the field. Furthermore, building on the results here, we would like to look into information-theoretic approaches to better understand how the emerging interpretable models could benefit social science, medical applications, and urban computing.

7.2.3 Event-driven Representation Learning

The success of machine learning algorithms generally depends on data representation. This is because different representations can entangle and hide the different explanatory factors of variation behind the data. Recently, representation learning has been proven effective in computer vision and natural language processing. We plan to investigate a few specific objectives: (1) develop new methods for representation learning driven by event prediction, (2) study feature and instance inter-dependencies in correlated tasks, and (3) reveal the relationship between events and factors (such as entity, locations, and organizations).
Bibliography


106


