Leveraging Heterogeneous Data Sources for Civil Unrest Prediction

Lu Meng, Rohini K. Srihari
Department of Computer Science and Engineering, State University of New York at Buffalo
lumeng@buffalo.edu, rohini@buffalo.edu

Introduction

Predicting significant societal events and generating early warnings is a challenging and critical problem since it involves multiple societal factors, including economics, politics, environment, and culture. Civil unrest prediction tasks have thus far relied on manually curated data sources and heuristic features determined by human expertise. Furthermore, current research focuses on machine learning approaches which do not effectively use heterogeneous data sources reflecting sequential data. In this paper, we propose a novel predictive framework which effectively exploits such data sources through LSTM networks. Extensive experiments have been conducted on 2 different datasets related to 2 countries to illustrate the effectiveness of our model.

Research Problem

Our goal is to forecast CU events counts by ingesting multiple data sources.

Suppose $X_t \in \mathbb{R}^n$ is a feature vector corresponding to a day t, where each entry $x_i \in \mathbb{R}^d$ of the vector contains the historical data from the previous d days, and $y_t + \Delta_t \in \{0, \dots, k\}$ refers to the category indicating the number of events on the day of $t + \Delta_t$. We are trying to produce f:

$$f : \mathbb{R}^n \to \{0, \dots, k\}, \text{ such that } y_t + \Delta_t = f(X_t)$$

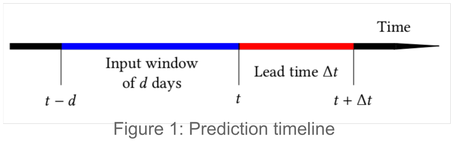


Figure 1: Prediction timeline

Data Sources

Historical Data

- Gold Standard Reports (GSR) data sets provided by IARPA for the Mercury Challenge
- The Armed Conflict Location & Event Data Project (ACLED) data

Figure 2: Monthly GSR events frequencies

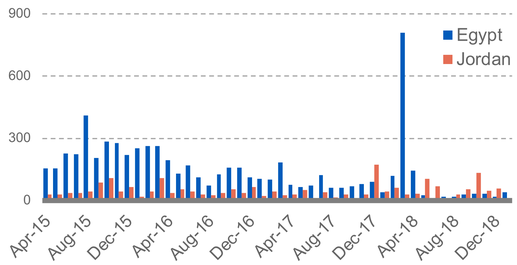
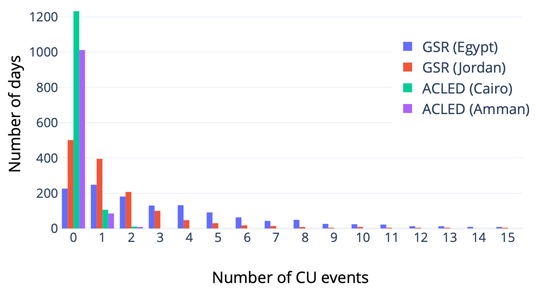


Figure 3: Histogram of number of events per day



Economic Indicators

- Commodity prices, unemployment rate, inflation rate, etc.

Table 1: Covariances between economic indicators and GSR event counts

Indicator	Egypt	Jordan	Indicator	Egypt	Jordan
Cotton	-9.83	-0.29	Gold	-2202.15	-228.08
Rice	-741.70	-1.82	Natural gas	-21.12	**7.22**
Wheat	-423.98	**56.12**	Iron ore	-431.48	**1.17**
Maize	**362.73**	**25.43**	Copper	-26373.92	**1470.69**
Sugar	-0.61	-0.24	Unemployment	**58.64**	-0.41

Social Media Data

- Political tweets from politicians and journalists.
- Daily volume and sentiment (percentage of angry posts) are then calculated

Methodology

Input layer takes in vector representations of indicator values within the input window. A series of 1 dimensional convolutional layers are added as well as the dropout layers. Output of the convolutional layer are input into a stack of bidirectional LSTM layers. Finally a softmax layer is used as the output layer for the classification task.

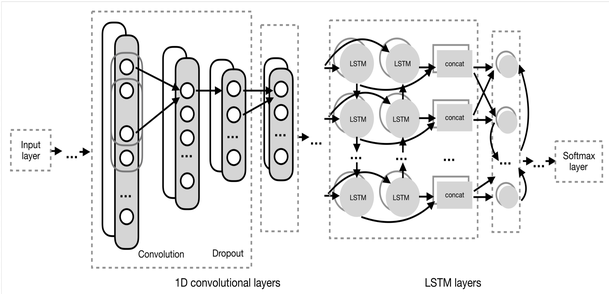


Figure 4: Framework of our model

Evaluation

Country level prediction

- From May 01, 2015 to January 08, 2019
- Ground truth data: GSR data for Egypt and Jordan
- Base rate model: ARIMA algorithm
- Performance evaluated according to Mercury Score (MS):

$$Quality\ Score = 1 - \frac{abs(Predicted - Actual)}{max(Predicted, Actual, 4)}$$

$$MS = 1{,}000{,}000 * QS$$

Table 2: Comparisons of Mercury Scores for each evaluation period.

	2018-08-01 2018-10-31	2018-11-01 2018-11-13	2018-11-01 2018-11-27	2018-11-01 2018-12-11	2018-11-01 2018-12-25	2018-11-01 2019-01-08
Egypt						
ARIMA	316479	269231	388889	390244	372727	362319
Cov-LSTM	798007	**846153**	787037	**829268**	**831818**	**829710**
rekcahd	**835015**	836152	**789545**	772238	802162	806661
valilenk	823034	788462	731481	762195	768182	778986
Jordan						
ARIMA	634114	**1000000**	**888889**	**758608**	**730986**	**699858**
Cov-LSTM	**658532**	647435	698765	680081	691060	699348
rekcahd	521188	768555	679934	571972	615693	647649
valilenk	378845	500000	515873	595238	612670	572817

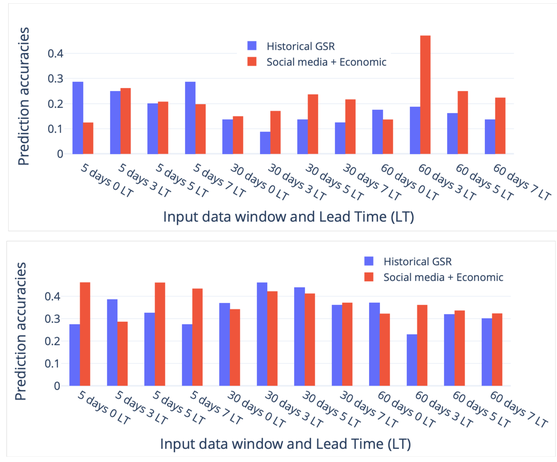


Figure 5: Prediction accuracies on GSR data under different system settings for Egypt and Jordan

City level prediction

- Ground truth data: ACLED data for Cairo and Amman
- Lack of data points
- One observation is that more false negative cases than at country level

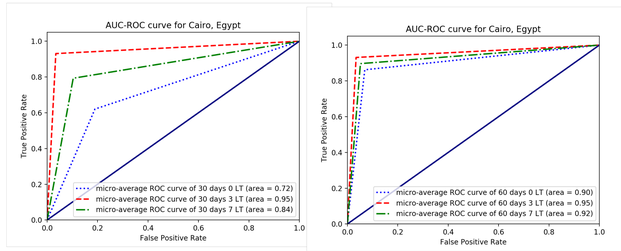


Figure 6: ROC curves of historical data based model on ACLED data for Cairo, Egypt.

Conclusions

We see promising results by including heterogenous data sources for predicting civil unrests and using LSTM models that effectively exploit sequential data. This work shows the possibility of leveraging existing data sets to provide predictions of civil unrest with sufficient lead time and granularity to be used in a deployed early warning system for effective resource allocation, safety and security planning and other decision making tasks.