Crystal Cube: Multidisciplinary Approach to Disruptive Events Prediction

Nathan H. Parrish[1], Anna L. Buczak[1,*], Jared T. Zook[1],
James P. Howard II[1], Brian J. Ellison[1], Benjamin D. Baugher[1]

[1] Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Rd, Laurel, MD 20723, USA
{nathan.parrish, anna.buczak, jared.zook, james.howard,
brian.ellison, benjamin.baugher}@jhuapl.edu

Abstract. The goal of Crystal Cube is to create an automated capability for the prediction of disruptive events. In this paper we present initial prediction results on six prediction categories previously shown to be of interest in the literature. In particular, we compare the performance of static classification models, often used in previous work for these prediction tasks, with a gated recurrent unit sequence model that has the ability to retain information over long periods of time for the classification of sequence data. Our results show that the sequence model is comparable in performance to the best performing static model (the random forest), and that more work is needed to classify highly dynamic prediction categories with high probability.

Keywords: prediction, disruptive events, gated recurrent unit, feature selection, multi-model analysis

1 Introduction

The goal of Crystal Cube is to create an automated capability for the prediction of disruptive events. The disruptive events we want to predict are wide-ranging and include armed conflict, insurgency, overthrow of dictators, economic collapse, failed states, and novel attacks on the US and other countries. Such a capability has broad interest to decision makers and leaders across a wide variety of domains including business, military and politics.

The ultimate goal of Crystal Cube is to predict a broad variety of types of disruptive events with high spatiotemporal resolution and maximum lead time. In this paper, we describe our initial approach and preliminary results. We develop models to predict six classes of events previously shown to be of interest by the research community: Domestic Political Crisis, Insurgency, International Crisis, Rebellion, Ethnic/Religious Vi-

[*] Corresponding Author

olence, and Irregular Leadership Change. We describe these categories in detail in section 3.1. In this initial work, we train models to predict each category of disruptive event on a country-month basis one month in advance.

One challenge in building predictive models for disruptive event prediction is that of input data design. Specifically, to train a classifier to predict the occurrence of disruptive events, a key decision is what features will be given to the model and at what time-lags. This decision becomes more challenging as the number of available input features grows. Previous studies have addressed this issue by one of two methods: either the features are chosen by some feature selection method [1], or the features are chosen based on an expert opinion of the important features for the given prediction category [2, 3]. Although these decisions address the issues of which features to include, the question of what time-lags to include in the feature set is often handled heuristically, with several months of lagged data being common.

One way to address the challenge of a decision on time-lags is to use a sequence modeling approach that is capable of retaining important information over time to help make decisions later in the sequence. Hidden Markov, linear state-space models that allow information to be passed over time through the incorporation of discrete valued hidden variables, have been used previously in event prediction [4]. More recently, recurrent neural networks and their extensions, gated recurrent units and long short-term memory units, have been gaining traction in sequence modeling problems due to their ability to retain hidden state information in a continuous valued memory variable. However, to the best of our knowledge, such models have yet to be applied to disruptive event prediction.

In this paper, we compare five different classification methods for predicting disruptive events: logistic regression, linear and radial basis function support vector machines, random forests, and gated recurrent units (GRU). GRU is a type of non-linear sequence model and, to the best of our knowledge, it has not been applied before to the problem of prediction of disruptive events. The paper is organized as follows: section 2 describes the related work, section 3 talks about the methods employed, including a description of the data we are using, section 4 describes the results, and section 5 contains our conclusions.

2 Related Work

Our work is related to other projects attempting to predict disruptive events. The Integrated Crises Early Warning System Project (ICEWS) uses the ICEWS coded event database augmented with macro-structural variables from various data sources like the World Bank to predict five of the six disruptive event categories that we predict in this paper. The ICEWS prediction categories include Domestic Political Crisis, Insurgency, International Crisis, Rebellion, and Ethnic/Religious Violence. The ICEWS data as well as the ground truth are available online [5].

Several prediction approaches have been applied to the ICEWS prediction categories. Montgomery, et al. [2] used ensemble Bayesian model averaging (EBMA) to fuse the forecasts of multiple classifiers. Arva, et al. [1] compared the performance of classification models using inputs derived from the ICEWS event database against another coded event database called the Global Database of Events, Language and Tone

(GDELT). They found that the inputs derived from GDELT provided as good or better performance than those from ICEWS. They additionally found that a combination of macro-structural variables and a subset of coded-event variables selected through a Bayesian model averaging approach, as opposed to all of the available input variables, was sufficient for accurate prediction. Neither of these studies considered sequential prediction models.

Beger, et al. [3] developed a prediction model for Irregular Leadership Change, a category not considered in the ICEWS project. They developed an ensemble-based, split-population duration model for the prediction problem. Each model within their ensemble was trained for a specific "theme" (e.g. public discontent or leadership characteristics). Each thematic model was a split-population duration regression that can be thought of as consisting of two components: a probability estimate of a countries belonging to either an ”at-risk of failure” class vs. ”not at-risk of failure” and then a regression conditional on this first estimate. Features for each theme were hand-selected from three different types of data sources: macro-structural, ICEWS coded event, and finally spatial variables for neighboring countries.

Qiao, et al. [4] developed a hidden Markov model (HMM) approach for predicting a custom truth category of social unrest events that they derived by looking for spikes of activitity in the GDELT event database. HMMs are a sequential model; however, they have been shown to provide lower performance for classification tasks than discriminative methods [16] like the GRU that we consider here.

3 Methods

3.1 Data Sources and Data Preprocessing

Crystal Cube uses open data sources to derive two types of input feature variables: global coded event data and social and economic meta variables or indicators. The difference in these two types of input variables and their utility for prediction of disruptive events is described in detail in [3].

The coded event features are extracted from the Global Database of Events, Language, and Tone (GDELT) [6]. GDELT is an open database that automatically documents societal activities around the world by applying natural language processing to contemporary news articles. Each entry into GDELT represents a unique news event and.GDELT provides basic contextual information about an event, and assigns it a code from the Conflict and Mediation Event Observations Event and Actor Codebook (CAMEO) [7]. CAMEO codes provide sensible categories to understand the nature of events. Crystal Cube uses as feature counts of events assigned to distinct CAMEO codes (e.g. 1122: accuse of human rights violations) lagged by one, two, and three months.

Social and economic features are derived from the World Development Indicators (WDI) [8] and Worldwide Governance Indicators (WGI) [9] that are compiled by the World Bank. The WDI data set includes over 1,000 indicators that estimate the level of development a country experiences year-to-year from a variety of perspectives. Examples of WDI include labor force participation rate by age, educational attainment by social class, amount of foreign aid received by a country, and national CO2 emissions.

WDI indicator values are available from 1960 through 2016. The WGI data set includes a small set of aggregates of expert opinions that describe the state of governance within a country in a given year. WGI estimates are split across six dimensions: Control of Corruption, Government Effectiveness, Political Stability and Absence of Violence/Terror, Regulatory Quality, Rule of Law, and Voice and Accountability. WGI indicator values are available for 1996, 1998, 2000, and 2002-2015.

The input data were preprocessed to generate monthly counts for different CAMEO codes and to fill in missing values. GDELT events were obtained from the GDELT 1.0 “reduced” event dataset. This is a preprocessed dataset which collapses the full GDELT database on "DATE+ACTOR1+ACTOR2+EVENTCODE" resulting in a single entry per event code per actor per day. These entries were then aggregated into monthly event code counts for each country by summing the number of events that occurred during the month with the country as either the source or target of the event. The GDELT reduced event dataset contains all events from January 1, 1979 through February 17, 2014. The WDI dataset had a large number of missing values. We removed WDI indicators containing 1000 or more missing values, and inferred missing values for the rest of the features by copying the most recent entry for that country.

As an additional set of features, we used a subset of the variables provided in the replication dataset of Beger, et al. [3]. The features that we included were those derived from ICEWS and those related to leadership characteristics within the country, (e.g., leader age and months in power).

We predict six categories of events that we call prediction categories or truth categories. The first five categories are derived from the “Events of Interest” ground truth dataset developed by the ICEWS project [5]. The five ICEWS prediction categories are: Domestic Political Crisis (an in-country political opposition to government not amounting to an insurgency or a rebellion), Insurgency (a coordinated effort to overthrow a government), International Crisis (escalating tensions between states / significant deployment of armed forces by one state in another’s territory), Rebellion (seeking independence from a government with ongoing organized, violent actions against it), and Ethnic/Religious Violence (violence between ethnic or religious groups that is not necessarily related to a government). If analysts concluded that one or more of these events occurred in a country over a specific timespan, they indicated it in the data on a month-by-month basis. Our sixth prediction category is Irregular Leadership Change; we derived this data from the truth set provided in [3]. The period over which we have ground truth for these six categories is March 2001 through March 2014.

3.2 Feature Selection

In total, we derive 1160 input features from the data sources described in section 3.1. In order to find a smaller set of features useful for predicting the truth categories, we perform feature selection by evaluating the information gain [14] and mutual information [15] between each input feature and each output category. Information gain and mutual information both provide a measure of the degree to which one random variable provides in predicting another, and are thus often used for feature selection. We perform feature selection for two reasons: previous studies have shown that a much smaller set of features is sufficient for disruptive event prediction [1], and additionally, some of our prediction models will not converge with such a large set of features.

For each of the six prediction categories, we computed the information gain, mutual information and the number of missing values. We removed all variables containing 1000 or more missing values, all the variables that had information gain and mutual information smaller than a certain threshold. For most of the categories the threshold of 0.03 was used for both information gain and mutual information. However, for Domestic Political Crisis and Irregular Leadership Change such a threshold would have resulted in no features being chosen. As such, thresholds 0.01 and 0.0002 were used for those two categories, respectively. The highest information gain for any feature was 0.004 for Irregular Leadership Change, making it evident that this would be the most difficult category to predict. Feature selection resulted in choosing 151 variables as inputs for Domestic Political Crisis (DPC), 118 as inputs for Ethnic/Religious Violence (ERV), 100 as inputs for Insurgency (INS), 135 as inputs for Rebellion (REB), 173 as inputs for International Conflict (IC), and 44 as inputs for Irregular Leadership Change (ILC).

3.3 Prediction Models

We compare five different prediction models: logistic regression, linear support vector machine (SVM), radial basis function support vector machine (RBF SVM), random forest (RF), and non-linear gated recurrent unit sequence model (GRU).

Logistic regression, support vector machine, and random forest are static classifiers that compute a prediction at time *t* based only on the features at that time. Such static classifiers have been used previously for disruptive event prediction [1, 2], and are described in detail in the texts [12, 13]. We focus our description in this section on the GRU model for disruptive event prediction.

The GRU is a sequential model designed for the prediction of sequence data. Figure 1 shows that, in contrast to static classification models, sequence classification models allow information from previous iterations to influence the predictions at the current time step through the transfer of latent variables. We believe that this ability to transmit information between time-steps is critical for the prediction of disruptive events as events can be influenced by sequences of previous events that occur over long-evolving time periods.

The GRU is defined by the following quantities that are computed at each time step as described in [11]: a memory unit $\boldsymbol{h}_t \in R^{d_h \times 1}$, a candidate memory unit $\widetilde{\boldsymbol{h}}_t \in R^{d_h \times 1}$, an update gating unit $\boldsymbol{z}_t \in [0,1]^{d_h \times 1}$, a reset gating unit $\boldsymbol{r}_t \in [0,1]^{d_x \times 1}$, and the input $\boldsymbol{x}_t \in R^{d_x \times 1}$. The following equations govern the interactions of these quantities:

$$\boldsymbol{h}_t = \boldsymbol{z}_t \odot \boldsymbol{h}_{t-1} + (1 - \boldsymbol{z}_t) \odot \widetilde{\boldsymbol{h}}_t,$$

$$\widetilde{\boldsymbol{h}}_t = \tanh\,(W^{(h,x)}\boldsymbol{x}_t + \boldsymbol{r}_t \odot U^{(h,x)}\boldsymbol{h}_{t-1}),$$

$$\begin{aligned} \boldsymbol{r}_t &= \sigma(W^{(r,x)}\boldsymbol{x}_t + U^{(r,h)}\boldsymbol{h}_{t-1}), \\ \boldsymbol{z}_t &= \sigma(W^{(z,x)}\boldsymbol{x}_t + U^{(z,h)}\boldsymbol{h}_{t-1}), \end{aligned}$$

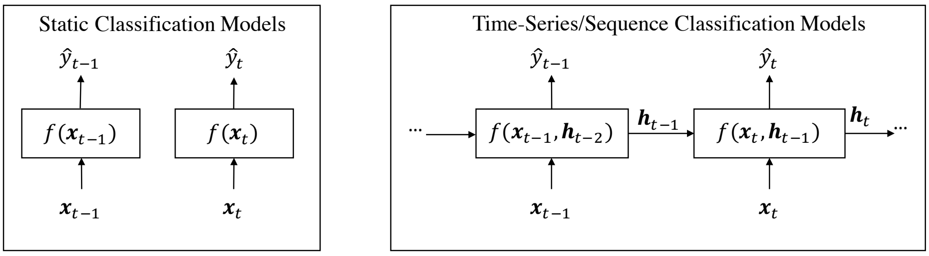

Fig. 1. Static classification models including logistic regression and support vector machines, make a prediction for time *t+1* based only on the features provided at time *t*. Time-series classification models make the decision for time *t+1* based on the features at time *t* and a hidden variable, $\boldsymbol{h}_t$, that carries information from previous time periods.

where $\odot$ is the element-wise matrix product, $\tanh(\cdot)$ is the element-wise hyperbolic tangent function, and $\sigma(\cdot)$ is the element-wise sigmoid function. The interaction between the memory units and the gating functions control how current and past information is stored and transferred and allow the GRU to retain information over long time periods. See Cho, et al. [10] or Chung, et al. [11] for more details. Finally, the prediction for time *t* is computed as a function of the current memory:

$$\hat{y}_t = \sigma(u^T \boldsymbol{h}_t + b).$$

4 Results

For each method (logit, linear SVM, RBF SVM, RF, GRU), we train six different prediction models, one for each prediction category (DPC, ERV, IC, INS, REB, ILC). The training data consists of the data from all 158 countries in our dataset from March 2001 – December 2011, and we test on data from January 2012 – March 2014. For each method, the input data for month *t*, model *k* consists of the features selected in our feature selection for the k^{th} prediction class as described in Section 3.2. Additionally, we train all the models with all the features to compare the results with and without feature selection.

Table 1 contains the area under the curve (AUC) for each of the machine learning methods for each of the six classes of events that we are predicting. All the methods have the results for the selected set of features. Only GRU, RFs, and linear SVM have results for all the features because logistic regression and RBF SVM failed to converge when using all the features.

The receiver operating characteristic (ROC) curve for each model is presented in Figure 2, with AUCs also printed in the legend. For the GRU, RF, and linear SVM, we plot only the better of the selected features / all features results based on which feature type gave the higher AUC for a given prediction category. From Table 1, we can see that this is the all features model for the GRU and RF and the selected features model for the linear SVM.

The RF and GRU are the best performing methods in terms of AUC. RFs perform the best in 4 out of 6 prediction categories, GRU performs the best in 1 out of 6 categories and in one category the AUC obtained by RFs and GRU is exactly the same (better than any other method). RFs consistently perform the best when all features are used. For the GRU the message is mixed (for 4 categories they perform better with all features, for one category they perform better with selected features, and for one category they perform the same). Linear SVM, RBF SVM and Logit perform the best when the feature selection is performed (they even sometimes fail to converge when all features are used).

Another trend that is notable in the results of Figure 2 and Table 1 is is that there is clear differentiation in how predictable the various truth categories are, with ethnic/religious violence, insurgenceny, and rebellion being highly predictable by some method, and domestic political crisis, international conflict, and irregular leadership change being moderately predictable by some method.

Figure 3 shows the ground truth for Nigeria and Pakistan, respectively for each of the six prediction categories. ERV, INS and REB, often stay at one level for a long time. DPC and IC are much more dynamic. The category that is the most difficult to predict is ILC as it encodes events that spike for only a single month.

Table 1. AUC for all prediction categories, all prediction models, and selected vs. all features. All features results are not available for the RBF support vector machine and the logit as these models failed to converge.

	Domestic Political Crisis	**Ethnic Religious Violence**	**Inter-national Conflict**	**Insurgency**	**Rebellion**	**Irregular Leadership Change**
All Features						
GRU	**0.82**	**0.98**	**0.86**	**0.97**	**0.99**	**0.79**
Lin SVM	**0.74**	**0.81**	**0.74**	**0.90**	**0.88**	**0.54**
Random Forest	**0.86**	**0.97**	**0.89**	**0.98**	**0.99**	**0.83**
Selected Features						
GRU	**0.80**	**0.98**	**0.89**	**0.93**	**0.93**	**0.76**
Lin SVM	**0.73**	**0.95**	**0.77**	**0.93**	**0.88**	**0.38**
Random Forest	**0.87**	**0.96**	**0.92**	**0.98**	**0.98**	**0.64**
RBF SVM	**0.74**	**0.83**	**0.89**	**0.94**	**0.94**	**0.52**
Logistic Regression	**0.78**	**0.95**	**0.75**	**0.93**	**0.90**	**0.72**

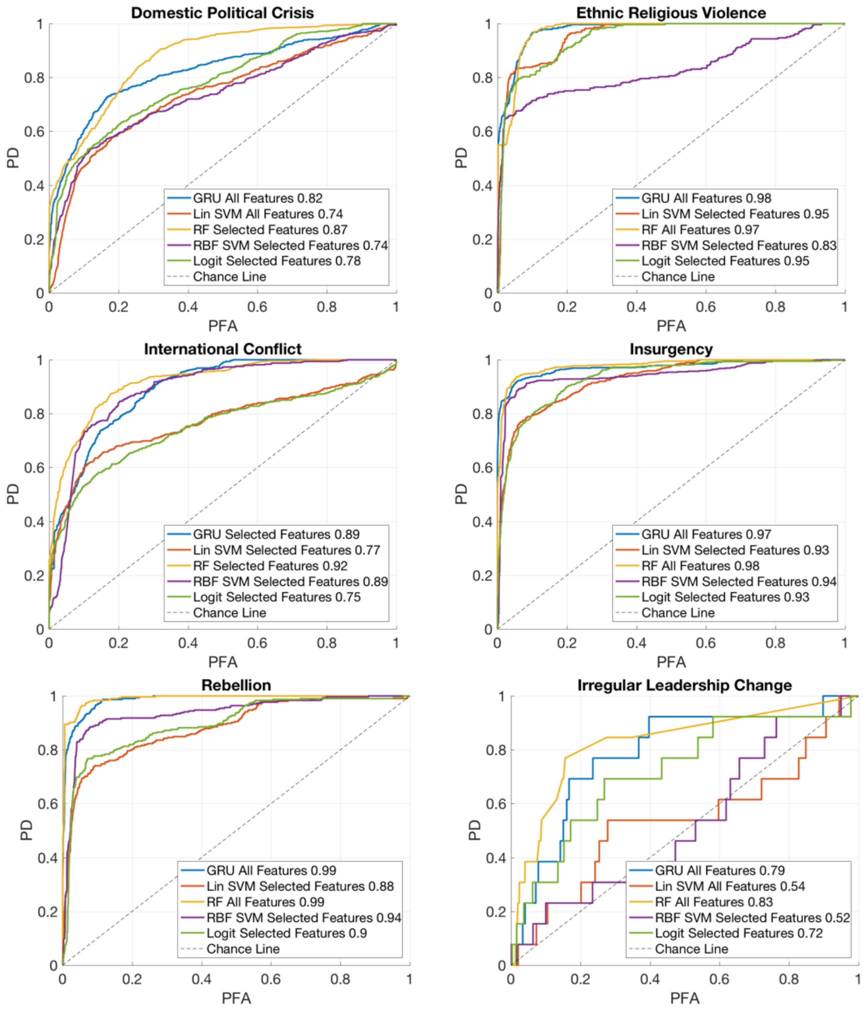

Fig. 2. ROC curves on the test data (Jan 2012 – March 2014) for the six different prediction categories. AUC values for each prediction model are printed in the legends for the model class.

Our models do not use the country name as one of the inputs. They also don't use lagged versions of the truth categories DPC, ERV, IC, INS, REB or ILC. If they did, we could get a higher AUC for relatively stable categories such as ERV, INS and REB.

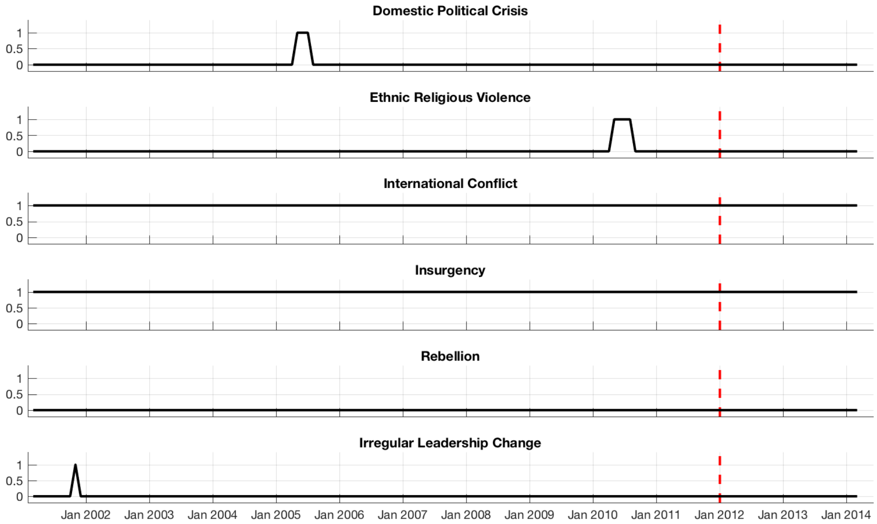

Fig. 3. Plot of the ground truth for Afghanistan. A zero indicates that the event did not occur in the given month and a one indicates that it did. The red horizontal line delineates the training period from the test period.

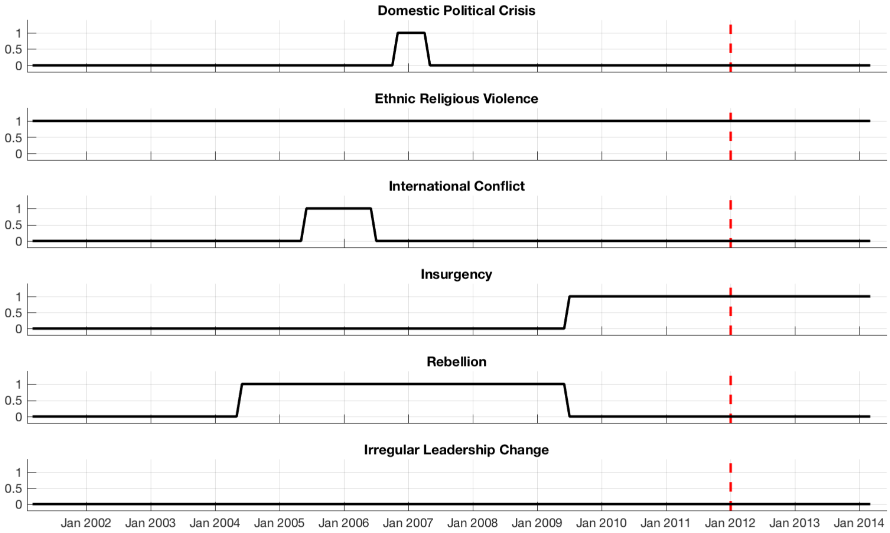

Fig. 4. Plot of the ground truth for Nigeria. A zero indicates that the event did not occur in the given month and a one indicates that it did. The red horizontal line delineates the training period from the test period.

5 Conclusions

Prediction of disruptive international political and security events is of great importance for several reasons. Economists and investment professionals would benefit from knowledge of what might happen in many regions in order to forecast how markets might react. Foreign policy makers might look to prediction in understanding how they might better engage with other nations and how U.S. policy might be adjusted. Furthermore, national security decision makers could be better informed in their decision process with foresight into the events in other nations. The deployment of military force, the enforcement of sanctions, and the preparation of market and currency disruption could better be prepared for if prior knowledge were more precisely understood.

This paper describes our initial approach and preliminary results for predicting six categories of disruptive events in the world (DPC, ERV, IC, INS, REB , ILC). Five methods (logit, linear SVM, RBF SVM, RF, GRU) were used to train the models to predict those disruptive events. RFs and GRUs were consistently better than other methods. In our future work we are planning to concentrate on predicting the categories that are the most difficult to predict (ILC, DPC, IC) due to their dynamic nature in individual countries. One approach could be to concentrate on predicting the change from the current status instead of predicting the value of the category for the next month.

The predictions were made at the country-level, and one month in advance. In the future we are planning to perform predictions at a higher spatiotemporal level (city or province) and more than a single month in advance.

In the present approach we used several open source data sets: GDELT, WDI and WGI. In the future we are also planning to use social media (e.g., Twitter) data sets in order to produce additional features for the prediction classifiers. We hope that these social media-based features will provide new and discriminative information for prediction that can complement our current input data sources.

References

1. Arva, B, Beieler, J., Fischer, B., Lara, G., Schrodt, P.A., Song, W., Sowell, M. and Stehle, S.: Improving Forecasts of International Events of Interest (2013)
2. Montgomery, J.M., Hollenbach, F.M., and Ward, M.D.: Improving Predictions Using Ensemble Bayesian Model Averaging. In: Political Analysis, Vol 20 (3), pp 271-291(2012)
3. Beger, A., Dorff C.L., Ward, D.: Irregular Leadership Changes in 2014: Forecasts using ensemble, split-population duration models. In: Int. Journal of Forecasting, Vol 32 (1), pp.98-111 (2016)
4. Qiao, F., Li, P., Zhang, X., Ding, Z., Cheng, J., and Wang, H.: Predicting Social Unrest Events with Hidden Markov Models Using GDELT. In: Discrete Dynamics in Nature and Society (2017)
5. Lustick, I., O'Brien, S., Shellman, S., Siedlecki, T., Ward, M.: ICEWS Events of Interest Ground Truth Data Set. Harvard Datavers, https://dataverse.harvard.edu/dataverse/icews (2015)
6. Leetaru, K.H., et al.: Global Database of Events, Language and Tone 1.0. The GDELT Project, https://www.gdeltproject.org (2017)

7. Schrodt, P.A.: CAMEO Conflict and Mediation Event Observations Event and Actor Codebook. Event Data Project, Pennsylvania State University Department of Computer Science (2012)
8. 2016 World Development Indicators. The World Bank, https:/data.worldbank.org/data-catalog/world-development-indicators (2016)
9. 2016 World Governance Indicators. The World Bank, https://data.worldbank.org/data-catalog/worldwide-governance-indicators (2016)
10. Cho, L. van Merrienboer, B. Bahdanau, D., and Bengio, Y.: On the Properties of Neural Machine Translation, Encoder-Decoder Approaches. In: Syntax, Semantics and Structure in Statistcal Translation, Vol 103 (2014)
11. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. NIPS Deep Learning Workshop, (2014)
12. Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning. Springer, New York (2001)
13. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
14. Kent, J.T.: Unformation Gain and a General Measure of Correlation. In: Biometrika, Vol 70 (1), pp 163-173 (1983)
15. Cover, T.M and Thomas, J.A.: Elements of Information Theory. John Wiley & Sons, New Jersey (2012)
16. Lafferty, J., McCallum, A. and Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: International Conference on Machine Learning (2001)