2020 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC) | 978-1-7281-8448-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/CYBERC49757.2020.00036

An Online Framework for Temporal Social Unrest Event Prediction Using News Stream

Qiao Fengcai
College of Advanced Interdisciplinary Studies
National University of Defense Technology
Changsha, China
fcqiao@nudt.edu.cn

Deng Jinsheng
College of Advanced Interdisciplinary Studies
National University of Defense Technology
Changsha, China
jsdeng@nudt.edu.cn

Wei Li
Troop 96766
Xinyang, China
kedaliwei2010@163.com

***Abstract*—Continuously assessing the risk of upcoming social unrest events and predict the likelihood of these events are of great importance. Thanks to the era of big data, people's understanding, experience, values, and ideology are mirrored in the organization of cyber space. In this paper, we propose an online prediction framework, using frequent subgraph patterns and hidden semi-Markov models (HSMMs). The feature called BoEAG (Bag-of-Event-Association-subGraph) is constructed based on frequents subgraph mining and the bag of word model. The new framework leverages the large scale digital history events captured from GDELT (Global Data on Events, Location, and Tone) to characterize the transitional process of the social unrest events' evolutionary stages, uncovering the underlying event development mechanics and formulates the social unrest event prediction as a sequence classification problem based on Bayes decision. Experimental results with data from Thailand demonstrate the effectiveness of the new framework, which outperforms the traditional HMM by 12.7% and the logistic regression 37.4%.**

***Index Terms*—frequent subgraph, graph mining, big data, algorithm**

I. Introduction

The social unrest events such as protests, strikes, demonstrations and occupy movements are important research focuses in the social computing area, which are common happenings in both democracies and authoritarian regimes [1]. Most social unrest events initially intended to be a demonstration to the public or the government. However, in many occasions they often escalate into general chaos, resulting in violent, riots, sabotage, and other forms of crime and social disorder. Take Thailand for an example, a series of political protests and three military coups happened between 1990 to 2015, resulting to the government being deposed, illustrating the power of the social unrest.

Traditionally, the research in the area of social unrest was based on static analysis from the macro-qualitative perspective by the political researchers. Fortunately, with the development of data science, especially the rise of big data, there are more and more data-driven approaches proposed on microscopic insight into possible social unrest events. Last century, most researchers conducted the prediction work using human-coded data, including WEIS [2] and COPDAB [3]. In the recent two decades, several small-scale vertical machine-readable datasets [4], [5] and large scale coded event data like ICEWS (Integrated Crisis Early Warning System) [6] and GDELT [7] appeared, fueling the development of computation methods for the analysis and prediction of social unrest.

Our previous work [8], [9] builded a hidden Markov models (HMMs) based framework to predict indicators associated with country instability. The framework used the temporal burst patterns in GDELT event streams as features to train the hidden Markov models. There are two shortcomings in that work. First, the temporal burst pattern is essentially a simple feature in the number of coded events which losses the interaction characteristics between event participants. Second, the probability of state residence time in the HMMs decreases exponentially with time, which is obviously not in line with the actual situation of social unrest events.

In response to the above shortcomings, we propose an online prediction framework in this paper, using frequent subgraph patterns and hidden semi-Markov models (HSMMs). The new framework also leverages the large scale digital history events captured from GDELT to characterize the transitional process of the social unrest events' evolutionary stages. Our proposed framework converts the GDELT event streams to frequent subgraph patterns for capturing interaction features better. In addition, the mechanism of HSMM guarantees the prediction model can explicitly learns the probability distribution of state residence time from the historical data. Eventually, the social unrest event prediction is formulated as a sequence classification problem using Bayes decision. More concretely, our main contributions in this updated paper are two pronged:

- First, we propose the BoEAG (Bag-of-Event-Association-subGraph) features to capture the characteristics of frequent patterns instead of the temporal burst patterns used in our previous work [9]. The original GDELT data within a certain time are represented as an event element association graph, from which the frequent subgraph patterns are mined. In the end, the BoEAG features are constructed like the classic BoW (Bag of Word) model [10] used in the text processing.
- Second, we propose a hidden semi-Markov model based framework which contains four major components: ground set extraction, BoEAG feature construction, HSMM training, and event prediction. The ground set

978-1-7281-8448-7/20/$31.00 ©2020 IEEE
DOI 10.1109/CyberC49757.2020.00036

contains social unrest events that are significant enough to garner more-than-usual real-time coverage in mainstream news reporting. The BoEAG features of the GDELT stream are taken as the observations. Then, two HSMM models are trained, with one for social unrest prone sequences and one for social unrest free sequences, after which new sequences' likelihoods are calculated and predictions are made by Bayes decision theory to specify the classification rule.

The paper is organized as follows: A coarse introduction of related work is provided in Section 2. Our HSMM based social unrest event prediction framework is presented in Section 3. In Section 4, extensive experiments to evaluate the performance of the new method are conducted and analyzed. The work is summarized and conclusions drawn in Section 5.

II. Related Work

Most recent social unrest event prediction techniques can be categorized into three types: planned event forecasting, classification based prediction and time series mining.

Planned event prediction methods don't need to mine patterns from the previous data. They are based on the hypothesis that protests that are larger will be more disruptive and communicate support for its cause better than smaller protests. Mobilizing large numbers of people is more likely to occur if a protest is organized and the time and place announced in advance [1], [11], [12]. For example, Basnet S et al. [13] used the GDELT data to propose a clustering method based on spatiotemporal k-dimensional structure trees to study the spatiotemporal distribution of conflict events in India in 2014.

Classification based prediction incorporates volume features and informative features such as semantic topics to train a classification model, and then predicts the occurrence of future events. Several classification methods are utilized such as random forest [14], support vector machines [15], logistic regression [16]–[18] and LASSO based logistic regression [19], [20]. Wang et al. [21] used the LSTM model combined with GDELT's event data to predict the number of conflicting events. Zhao et al. [22] used the multi-task learning of geographical spatial stratification, judging whether unrest events occurred on the specified date. Wu et al. [23] used the "Protest Participation Theory" proposed in the field of political science, combined with the SVM support vector machine model to conduct early warning research on social unrest events. Deng et al. [24] extracted and learned graph representations from historical event documents. By employing the hidden word graph features, the model predicts the occurrence of future events and identifies sequences of dynamic graphs as event context.

Time series based mining uses temporal correlation of relevant features such as tweet volume by adopting appropriate approaches. For example, Achrekar et at. [25] used autoregressive modeling to predict flu trends using twitter data. Radinsky et al. [26] utilized NYT news articles from 1986 to 2007 to build event chain and identify significant increases in the likelihood of disease outbreaks, deaths, and riots in advance of the occurrence of these events in the world.

III. HSMMs-based Social Unrest Events Prediction

A. Framework

Proactive reaction to social unrest events is at first glance closely coupled with social unrest event detection: an unrest event needs to be detected before the government can react to it. However, the fact is that not the detection result but the eruption of a social unrest event is the kind of event that should be primarily avoided, which makes a big difference. Hence, it goes without saying that efficient proactive handling of social unrest events requires the prediction of the future level of social unrest, to judge whether the current situation bears the risk of an unrest event or not. The evolutionary stages of the social unrest event can not be directly observed. However, the stages have been explicitly coded more or less on the Internet. The basic assumption of our approach is that the eruption of social unrest events can be identified by frequent subgraph patterns of the event sequence prior to the happening time point using HSMMs.

The evolutionary stages of the social unrest event can not be directly observed. However, the stages have been explicitly coded more or less on the Internet. The basic assumption of our approach is that the eruption of social unrest events can be identified by frequent subgraph patterns of the event sequence prior to the happening time point using HSMMs. If a prediction is performed at time t, we would like to know whether a social unrest event will occur or not between time $t + \Delta t_l$ to $t + \Delta t_l + \Delta t_p$.

Our prediction task will resolve around predicting significant social unrest events on the country level and considering that country alone. To accurately predict social unrest events it is crucial to be able to characterize these events' underlying stage before the occurrence by utilizing relevant GDELT event records observations. We propose a hidden semi-Markov model based framework to characterize the underlying development of these events. Fig. 1 gives the proposed HSMMs-based social unrest event prediction framework, which contains four major components: ground set extraction, BoEAG feature construction, HSMM training, and event prediction. The ground set extraction and event prediction are the same as [9], not tired in words here.

Formally, denote ER as a basic GDELT event record. $ER("columnName")$ means the value of a specified column in a record. Denote $D = \{ER_{c,t}\}_{c\in\Omega,t\in\Gamma}$ as a collection of GDELT event record data split into different countries Ω in time period Γ. The country c and the day t can be filtered by $ER(ActionGeo\_CountryCode)$ and $ER(SQLDATE)$ respectively. Since event records ER are being added daily by the hundreds or thousands to the GDELT event table, we aggregate those event records by day, defined as $DAER_{c,t}$, meaning the daily aggregated event record on the day t in country c. Then a sequence of $DAER$s is defined as $s =$

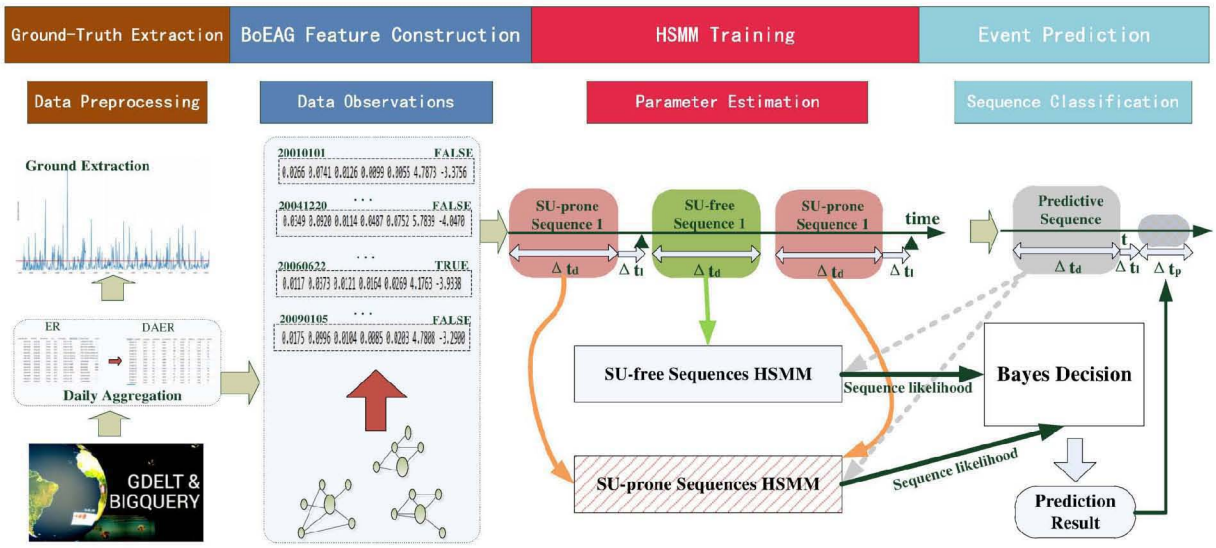

Fig. 1. The proposed HSMMs-based social unrest event prediction framework: Two HSMMs are trained, with one for SU-prone sequences and one for SU-free sequences. SU-prone sequences consist of observations (BoEAG feature) within a time window of length Δt_d preceding a social unrest event(▲) by lead time Δt_l. SU-free sequences consist of observations at times when no social unrest event was imminent. t is the time the prediction performed at. Δt_p is the prediction period.

$\{DAER_{c,t}\}_{t \in T \subseteq \Gamma}$, which contains all the daily aggregated event records in country c in the time period $T \subseteq \Gamma$.

B. BoEAG Feature Construction

The Bag-of-Event-Association-subGraph feature (BoEAG) is constructed from frequent subgraphs and the bag of word model. The original GDELT data within a certain time are first represented as a big single event element association graph. Then, the frequent subgraph patterns are mined from the big single graph. In the end, the BoEAG features are constructed like the classic BoW (Bag of Word) model.

The event element association graph draws on the SUBDUE system [27] which analyzed aviation safety events using graph mining. The system converts a series of aviation safety related event records into graph data for processing. The node labels represent the aviation safety event id and the attribute value. The edge labels represent the attribute name (such as location, time, flight altitude) and the relationship between events. For example, "near_to" relationship means that the distance between the two accidents occurred is within 200km.

Bag of words model is a feature vectorization method commonly used in the field of text retrieval and text classification. In this paper, BoEAG feature construction is similar to BoW. The collection of GDELT event element association graphs aggregated by day corresponds to the corpus in the BoW model. Each event element association graph corresponds to a document and each frequent subgraph corresponds to a word in the BoW model. The $tf - idf$ weight of the frequent subgraph s of the $i-th$ event element association graph can be calculated by the following Formula 1:

$$\begin{cases} tfidf(s,i,D) = tf(s,i) \times idf(s,D) \\ tf(s,j) = log(1 + f_{s,i}) \\ idf(s,D) = log(\frac{N}{1+n_s}) \end{cases} \quad (1)$$

where $f_{s,i}$ denotes the frequency of subgraph s in the event association graph i. This value can be directly obtained through the single graph frequent subgraph mining algorithm SSIGRAM proposed in our previous work [28]. N denotes the number of event association graphs, that is, the time span of the data set in days; n_s is the number of event association graphs that contain subgraphs s.

Algorithm 1 gives the process of BoEAG feature construction illustrated above. The input of the algorithm including three parameters: the original GDELT event records, such as a set of event records within a certain period of time in a certain country, the support threshold and the maximum number of subgraphs. The output is the BoEAG feature vector set . Lines 4 to 19 of the algorithm construct event association graphs. Lines 20-22 use the SSIGRAM algorithm for single large graphs for frequent subgraph mining. The maximum number of subgraphs N_{max} is to return the maximum number of subgraphs. That is, when the total number of frequent subgraphs found during the mining process reaches N_{max}, it will stop iterating and arrange all subgraphs in descending order of frequency. Line 24 obtains the standard adjacency matrix coding sequence of each subgraph and uses it as the "Word". Line 25 calculates the $tf - idf$ feature vector

corresponding to each event association graph according to the Formula 1 .

Algorithm 1 The algorithm of BoEAG feature construction

Input: original event records ER, support threshold τ, maximum subgraphs returned N_{max}
Output: BoEAG feature set X_{set}

$EAG_{set} \leftarrow \varnothing$ /*The set of event association graphs*/
$SubG_{set} \leftarrow \varnothing$ /*The set subgraphs*/
$X_{set} \leftarrow \varnothing$
$DAER_{list} \leftarrow ER$: event records aggregated by day
for $DAER_t$ in $DAER_{list}$ **do** /*All the event records at date t*/
    $EAG_t \leftarrow \varnothing$ /*All the event association graphs at date t*/
    for e_i in $DAER_t$ **do**
        if e_i isn't be traversed **then**
        $EAG_t \leftarrow$ constructing the graph unit of event e_i
        for e_j in $DAER_t$ **do**
            if e_j isn't be traversed **then**
                $EAG_t \leftarrow$ constructing the graph unit of event e_j
                if e_i and e_j contain at least one identical participant **then**
                $EAG_t \leftarrow$ generating "relate_to" edge between e_i and e_j
            end if
        end for
    end for
    $EAG_{set} \leftarrow EAG_t$
end for
for EAG_t in EAG_{set} **do**
    $SubG_t \leftarrow$ SSIGRAM(EAG_t,τ,N_{max}) /*Mining frequent subgraphs using the SSIGRAM algorithm*/
    $SubG_{set}.add(SubG_t)$
end for
Representing each subgraph in $SubG_{set}$ as its standard adjacency matrix (CAM) coding sequence. [a]
$X_{set} \xleftarrow{TF-IDF} SubG_{set}$ /*Calculating feature set using Formula 1*/
return X_{set}

[a]For details of standard adjacency matrix please refer to [28]

C. HSMM Training

Usually, the social unrest event has a series of evolutionary stages, through a longer or shorter life cycle, meaning that it is usually not a sudden outbreak. Typical stages in the events' life cycle often include: appeal, accusation, refuse, escalation and eruption.

In the traditional HMM model, the state residence time probability $P_i(d)$ shows an exponential downward trend with the number of residence time units [29], which is obviously not consistent with the state residence time of many application scenarios in the real world, especially the social unrest events. In order to improve this shortcoming, the state residence time probability distribution can be explicitly introduced into the HMM model, so that it can automatically learn the probability distribution of the state residence time from historical data. This is the original intention of the hidden semi-Markov model.

Let $S = \{s_i\}$ denote the set of latent states, $1 \leq i \leq N$. Let $\pi = [\pi_i]$ denote the vector of initial state probabilities. Given a sequence of the above BoEAG feature observations O, a standard continuous HSMM can be defined as $\lambda = (\pi, A, B, P)$, where the initial state probability π and output matrix B have the same meaning as HMM, while the state transition matrix A is defined as:

$$a_{ij} = P(S_{t+d} = s_j | S_t = s_i), \quad 1 \leq i, j \leq N \tag{2}$$

This paper considers the discrete time probability, that is, the state residence time can only be an integer multiple of the residence time unit e.g. day. Let D represent the maximum possible residence time, then P can be denoted as a residence time probability matrix of $N \times D$, whose element value p_{id} represents the probability of the state s_i lasting d time units:

$$p_i(d) = P(d | S_t = s_i), \quad 1 \leq i \leq N, 1 \leq d \leq D \tag{3}$$

Given an observation sequence consisting of L days' BoEAG feature vector set $O = (\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_L)^T$. The goal of hidden semi-Markov model training is to optimize the model parameters π, A, B and P, so that the likelihood of the model generating sequence O is maximized. Given the HSMM model $\lambda = (\pi, A, B, P)$, the sequence likelihood of the observation sequence O is defined as:

$$P(O|\lambda) = \sum_{\mathbf{s}} \pi_1 b_{s_1}(\mathbf{o}_1) \Pi_{t=2}^{N} P(S_t = s_t | S_{t-1} = s_{t-1}) b_{s_t}(\mathbf{o}_t), \tag{4}$$

Where $\mathbf{s} = [s_t]$ represents the hidden state sequence with length N. Similarly as the traditional HMM, the sum over $\mathbf{s}$ can also be calculated by the forward-backward algorithm proposed in [30]. The difference is that the state residence time needs to be explicitly added during the derivation process. Define $\alpha_t(j)$ as the forward variable, which means the probability of ending at the hidden state j at time t, given observation sequence $O = (\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_L)^T$. The backward variable is defined as $\beta_t(i)$, which means the probability of starting at the hidden state i at time t, given observation sequence $O = (\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_L)^T$.

Thus, there are 4 parameters to be estimated for the model training, including initial probability distribution π, state transition probability a_{ij}, observed probability density function $b_i(\mathbf{o}_t)$, state residence time probability density function $p_j(d)$. π and a_{ij} can be calculated directly. $b_i(\mathbf{o}_t)$ and $p_j(d)$ need to specify the description form of probability density function in advance. We use multivariate mixed Gaussian probability density to describe the probability density of observations $b_i(\mathbf{o}_t)$:

$$b_i(\mathbf{o}_t) = \sum_{m=1}^{M} c_{im} \mathcal{N}(\mathbf{o}_t; \mu_{im}, \mathbf{U}_{im}), \tag{5}$$

where M represents the number of mixed Gaussian elements. c_{im} is the weight of the m mixed Gaussian elements in the state i, and $\sum_{m=1}^{M} c_{im} = 1$. μ_{im} and $\mathbf{U}_{im}$ are the mean and variance of the $i-th$ Gaussian element respectively.

We use a single Gaussian distribution to describe the probability density of state residence time $p_j(d)$,

$$p_i(d) = \mathcal{N}(d; m_i, \sigma_i^2), \tag{6}$$

where m_i and σ_i^2 are the mean and variance, respectively.

Denote the variable $\xi_t(i,j)$ as the probability of transferring from state i to state j after residing in d time units at the time t. Given the observation sequence O and the model parameters λ, then

$$\xi_t(i,j) = P(S_t = i, S_{t+d} = j|O, \lambda). \tag{7}$$

Given the definitions of the forward variable and backward variable, $\xi_t(i,j)$ can be calculated as:

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}\sum_{d=1}^{t}\beta_{t+d}(j)p_j(d)\prod_{s=t+1}^{t+d}b_j(o_s)}{\beta_0}. \tag{8}$$

So far, the parameter estimation can be achieved by the Expectation Maximization (EM) algorithm, also known as the Baum-Welch algorithm in HMM [30]. The E step of the EM algorithm is to construct a Q function, and then maximize the Q function in the M step. Thus, we can obtain the re-estimated model parameters π, a_{ij}, $b_i(\mathbf{o}_t)$ and $p_j(d)$. Then the process iterates continuously until the parameters converge or the maximum number of iterations is reached, formulated as:

$$\hat{\lambda} = \underset{\lambda}{argmax}\, P(O|\lambda). \tag{9}$$

As the Ground-Truth contains multiple positive samples and negative samples, we need to use multiple sets of observation data to train the model. Denote $O = [O^{(1)}, O^{(2)}, ..., O^{(k)}, ..., O^{(K)}]$ as the training data containing K observation sequences. All observation sequences have the same length L. We assume that each observation sequence is independent with each other. $P(O|\lambda)$ represents the probability of the combination of observation sequences under a given model, then

$$P(O|\lambda) = \prod_{k=1}^{K} P(O^{(k)}|\lambda). \tag{10}$$

Finally, we trained two HSMMs based on two corresponding set of sequences, one set from sequences prior to the positive 7-day stretches minus the lead time period and the other negative. Thus, one model characterizes the evolution process leading to a social unrest event, while the other one characterizes the process that does not lead to a social unrest event.

IV. Experiment Evaluation

A. Experiment Design

Dataset: The dataset used is the same as [8]. Our goal in this paper is to predict the overall level of protest using GDELT, and our focus country is Thailand. As mentioned above, GDELT uses the CAMEO coding system [31], where root event code 14 means protest. The actual training data was from April 1, 2001 to December 31, 2013, and the test data January 1, 2014 to February 29, 2016. About 11.5% of 7-day stretches are labeled positive, distributed mostly evenly among the countries. The whole training and testing period include 5448 days, 778 weeks. The training period includes 666 7-day stretches while the testing period 112. The Number of positive 7-day stretches in training and testing data are 95 and 12 respectively.

Comparison Methods: As a comparison, three methods are selected in this paper. One is the traditional hidden Markov model HMM. Except that there is no explicit state residence time probability distribution estimation during the model training process, the remaining steps are the same as the HSMM method. The second is *Logistic* regression method. Two *Logistic* regression model are trained and sequence classification is conducted based on this. The third is *baseline* which does not train any model. It directly uses the probability of protest event records in a country in history as the future social unrest events' probability.

Performance Metrics: We quantify the success of the proposed predictive mechanism and comparison methods based on their balanced accuracy. Let $T_{ct} \in \{0,1\}$, $P_{ct} \in \{0,1\}$ respectively denote whether a significant social unrest event occurs in country c during the days $t-3, t-2, t-1, t, t+1, t+2, t+3$ and whether we predict there to be one. The true positive rate (TPR) is the fraction of positive instances ($T_{ct} = 1$) correctly predicted to be positive ($P_{ct} = 1$) and the true negative rate (TNR) is the fraction of negative instances predicted negative. The balanced accuracy ($BACC$) is the unweighted average of these:

$$BACC = \frac{TPR + TNR}{2}. \tag{11}$$

Parameter Settings: In the extraction stage of Ground-Truth, the threshold value of θ is set to 2.3. This value is approximately equal to the 90% quantile of the standard exponential distribution, that is, approximately 10% of the 7-day time windows in the Ground-Truth will be marked positive.

In the BoEAG feature extraction stage, the maximum number of returned frequent subgraphs N_{max} is set to 10000. The *Logistic* regression has one parameter: the iteration convergence threshold, which is set to 10^{-6} in the experiment. The *baseline* method does not require any parameter values to be set in advance. The HMM model and the HSMM model both have 6 parameters need to be set, including the hidden state number N, the number of mixed Gaussian elements used in the estimation of the probability density of the observation value M, the prediction interval Δt_p, the lead time Δt_l, the prediction data time window Δt_d and the likelihood threshold ε. The final value details are shown in Table I.

TABLE I
ADJUSTABLE PARAMETERS Δt_l, Δt_p AND ε

Country	HMM			HSMM		
	Δt_l	Δt_p	ε	Δt_l	Δt_p	ε
Thailand	1	10	0.2	1	10	0.1

B. Event Prediction Results

Table II gives the balanced accuracy BACC values of the hidden semi-Markov model HSMM, the traditional hidden Markov model HMM, the logistic regression and the baseline method on the test set. Based on the BoEAG feature pattern, it can be seen that in the test data sets, the performance of the prediction method based on the hidden semi-Markov model proposed in this paper is the best, which shows that the HSMM model can indeed better model the characteristics of mass protest events due to explicitly considering the residence time of the event development evolution stage. The performance of the HMM model is the second, followed by the logistic regression, and the baseline performs the worst, which is basically random guessing.

TABLE II
THE BACC VALUE OF EACH METHOD.

	HSMM	HMM	Logistic	Baseline
BoEAG Pattern	**95.9%**	85.1%	69.8%	52.8%
Burstness Pattern	87.1%	86.1%	67.8%	53.1%

In addition, comparing each method's performance based on the BoEAG pattern and the temporal burstness pattern used in our previous work [9], we can see that the BoEAG pattern constructed from frequent subgraphs can better model the stages of social unrest events, as the BACC values of HSMM, HSMM and Logistic all improve when the BoEAG patterns are used. This is because the BoEAG pattern considers both temporal burstness and the interaction between event participants.

By adjusting the likelihood ratio threshold ε, a series of correspondences between the true positive rate TPR and the false positive rate FPR can be obtained, and then ROC analysis can be performed for each method. Fig. 2 shows the ROC curve of the three methods of HSMM model, HMM model and Logistic regression. The larger the area under the curve (AUC) under the ROC curve, the better the prediction performance of the model. Obviously, among the three methods shown, the AUC area of the hidden semi-Markov model HSMM is the largest on each test set, and its performance is the best among the methods.

C. Sensitivity Analysis on Δt_l and Δt_d

Although the model parameters are fixed on the training set by 10-fold cross-validation, it is still necessary to investigate the performance of the prediction model at different leading time Δt_l and prediction time window Δt_d, which also has guiding significance to the actual application model.

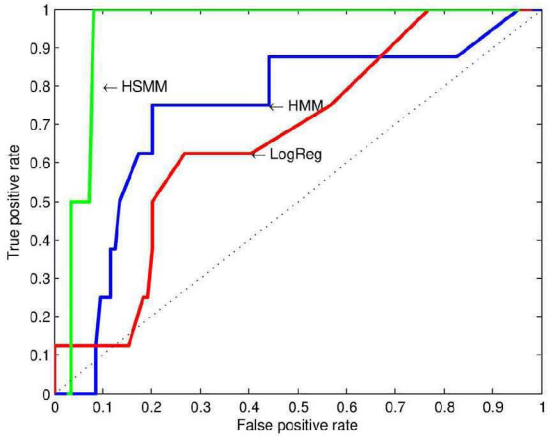

Fig. 2. ROC curves for the compared prediction models: HSMM, HMM, Logistic.

Fig. 3 shows the trend of the prediction performance of the HSMM model on each test set with Δt_l and Δt_d. The leading time Δt_l is 1 day to 10 days, and the value of Δt_d is 10 days, 20 days, and 30 days. Two phenomena can be found: First, as the leading time Δt_l increases, the overall prediction accuracy of the model decreases. In most cases, when $\Delta t_l = 1$, the BACC value is the highest. This is consistent with our common sense, that is, the closer the observation data is to the time point of the event, the more accurate the event can be predicted in the future. Second, the performance of the model is not necessarily related to the length of time windows of the observation sequence data used. It is not that the longer the observation sequence used, the higher the prediction accuracy, and the more data, the more interference. Given the trained prediction model and the lead time parameters, different test sets require different time windows for prediction data to achieve optimal prediction accuracy.

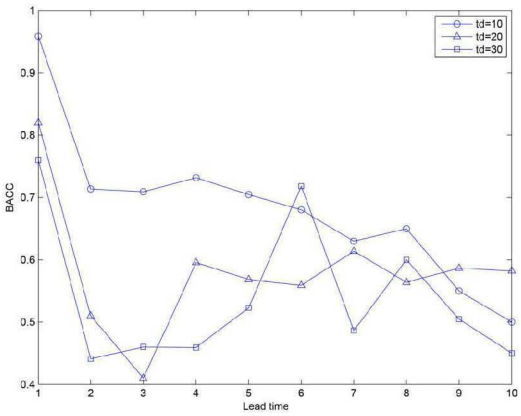

Fig. 3. Sensitivity analysis on lead times Δt_l and data window size Δt_d.

CONCLUSTION

This paper presents a hidden semi-Markov models based framework for leveraging large scale digital history coded events captured from GDELT to utilize the frequent subgraph patterns mined from the GDELT event streams to uncover the underlying event evolution mechanics and formulate the social unrest event prediction as a sequence classification problem. Extensive empirical testing with data from Thailand in the Southeast Asia demonstrated the effectiveness of this framework by comparing it with traditional HMM, the logistic regression model and the baseline model. It shows that the GDELT dataset do reflect some useful precursor indicators that reveal the causes or evolution of future events.

We plan to conduct our future work in the following three aspects. First, we plan to introduce a multi-level prediction mechanism to our framework, such as city level or province level. Second, in GDELT 2.0, event mention details and global knowledge graphs [32] are also provided real-timely, which can bring us with detail insights to the events. More machine learning and deep learning methods like the graph neural networks [33] can be developed with more events' elements. Third, the prediction framework may be improved by distinguishing widespread news coverage from localized coverage.

REFERENCES

[1] S. Muthiah, B. Huang, J. Arredondo, D. Mares, L. Getoor, G. Katz, and N. Ramakrishnan, "Planned protest modeling in news and social media." in *AAAI*, 2015, pp. 3920–3927.

[2] C. A. McClelland, "World-event-interaction-survey: A research project on the theory and measurement of international interaction and transaction," *University of Southern California*, 1967.

[3] E. E. Azar, "The conflict and peace data bank (copdab) project," *Journal of Conflict Resolution*, vol. 24, no. 1, pp. 143–152, 1980.

[4] D. Bond, J. C. Jenkins, C. L. Taylor, and K. Schock, "Mapping mass political conflict and civil society issues and prospects for the automated development of event data," *Journal of Conflict Resolution*, vol. 41, no. 4, pp. 553–579, 1997.

[5] S. P. Orien, "Crisis early warning and decision support: Contemporary approaches and thoughts on future research," *International Studies Review*, vol. 12, no. 1, pp. 87–104, 2010.

[6] B. Kettler and M. Hoffman, "Lessons learned in instability modeling, forecasting, and mitigation from the darpa integrated crisis early warning system (icews) program," in *2nd International Conference on Cross-Cultural Decision Making: Focus*, 2012.

[7] K. Leetaru and P. A. Schrodt, "Gdelt: Global data on events, location, and tone, 1979–2012," in *ISA Annual Convention*, vol. 2, no. 4. Citeseer, 2013.

[8] F. Qiao and K. Chen, "Predicting protest events with hidden markov models," in *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2016 International Conference on*. IEEE, 2016, pp. 109–114.

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

[10] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2010, pp. 841–842.

[11] S. Muthiah, "Forecasting protests by detecting future time mentions in news and social media," 2014.

[12] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz *et al.*, "'beating the news' with embers: forecasting civil unrest using open source indicators," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1799–1808.

[13] S. Basnet, L.-K. Soh, A. Samal, and D. Joshi, "Analysis of multifactorial social unrest events with spatio-temporal k-dimensional tree-based dbscan," in *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Analytics for Local Events and News*, 2018, pp. 1–10.

[14] N. Kallus, "Predicting crowd behavior with big public data," in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 625–630.

[15] J. Ritterman, M. Osborne, and E. Klein, "Using prediction markets and twitter to predict a swine flu pandemic," in *1st international workshop on mining social media*, vol. 9. ac. uk/miles/papers/swine09. pdf (accessed 26 August 2015), 2009, pp. 9–17.

[16] R. Compton, C. Lee, J. Xu, L. Artieda-Moncada, T.-C. Lu, L. De Silva, and M. Macy, "Using publicly visible social media to build detailed forecasts of civil unrest," *Security Informatics*, vol. 3, no. 1, pp. 1–10, 2014.

[17] F. Qiao and H. Wang, "Computational approach to detecting and predicting occupy protest events," in *2015 International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI)*. IEEE, 2015, pp. 94–97.

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

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

[20] G. Korkmaz, J. Cadena, C. J. Kuhlman, A. Marathe, A. Vullikanti, and N. Ramakrishnan, "Combining heterogeneous data sources for civil unrest forecasting," pp. 258–265, 2015.

[21] X. Wang, H. Chen, Z. Li, and Z. Zhao, "Unrest news amount prediction with context-aware attention lstm," in *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2018, pp. 369–377.

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

[23] C. Wu and M. S. Gerber, "Forecasting civil unrest using social media and protest participation theory," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 1, pp. 82–94, 2017.

[24] S. Deng, H. Rangwala, and Y. Ning, "Learning dynamic context graphs for predicting social events," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2019, pp. 1007–1016.

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

[26] K. Radinsky and E. Horvitz, "Mining the web to predict future events," in *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013, pp. 255–264.

[27] N. S. Ketkar, L. B. Holder, and D. J. Cook, "Subdue: Compression-based frequent pattern discovery in graph data," in *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*. ACM, 2005, pp. 71–76.

[28] F. Qiao, X. Zhang, P. Li, Z. Ding, S. Jia, and H. Wang, "A parallel approach for frequent subgraph mining in a single large graph using spark," *Applied Sciences*, vol. 8, no. 2, p. 230, 2018.

[29] , "ŁŁŁ [d]," Ph.D. dissertation, , 2011.

[30] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[31] D. J. Gerner, P. A. Schrodt, O. Yilmaz, and R. Abu-Jabr, "Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions," *International Studies Association, New Orleans*, 2002.

[32] "Gdelt 2.0: Our global world in realtime," http://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/.

[33] T. Kipf *et al.*, "Deep learning with graph-structured representations," Ph.D. dissertation, 2020.