Chapter 6

Social Media as Human Sensors for Forecasting Civil Protests

6.1 Introduction

Civil unrest or civil disobedience is referred as a social instability and protest movements at the National and International level primarily against the government and policymakers [186]. Civil unrest can be both non-violent demonstrations or strikes as well as violent riots. The reason behind large-scale civil unrest is mainly discontent in the society due to the poor social and economic conditions [187]. The Arab Spring democratic uprising which originated in Tunisia in December 2010\(^1\) and then propagated across various countries in the Arab world in the year 2011 is an example of intense civil unrest and disorder\(^2\). Similarly, in a recent incident (February 1, 2017) a violent protest took place at Berkeley University, California causing 100,000 USD worth of damage to the university campus\(^3\). Due to high reachability and popularity of social media websites worldwide, organizations use these websites for planning and mobilizing events for protests and public demonstrations [188]. The study of civil unrest reveals that now most of the protests are planned and mobilized in much advance [187] [189]. Traditionally, newspapers have been used as primary sources for such analysis and prediction. However, the speed and flexibility of publication on social media platforms gained the attention of various organizations for planning and making announcements of various protests, strikes, public demonstrations, and riots. Organizing such large-scale civil protests requires planning and mobilization, and it is seen that the due to its immense popularity and wide reachability, Twitter is being used as a platform for sharing information about civil protest events and mobilize them [187] [190]. Figure 6.1 shows some concrete examples of tweets published for planning and mobilization of several civil protests and unrest related events. Due to the presence of public posts about mobilization, an early prediction of such events can be done by applying OSSMInt on Twitter data. In countries like USA, India, and Australia where protests are legal, early detection or forecasting of such events is valuable for government, tourism and law enforcement agencies. For example, it can help police deployment in those areas to maintain law and order and prevent violence. Similarly, it can help the government to deploy local civil force to retain the route traffic and prevent the interactions between protesters and bystanders [187].

We conduct a literature survey in the area of mining social media for prediction of civil unrest events. Based on our review, we find that while there has been a lot of work done in the area of general-interest event forecast; the prediction of civil unrest and protest events has recently gained the attention of researchers

\(^{1}\)http://middleeast.about.com/od/humanrightsdemocracy/tp/Arab-Spring-Uprisings.htm \\
\(^{2}\)http://www.npr.org/2011/12/17/143897126/the-arab-spring-a-year-of-revolution \\
6.1 Introduction

We discuss a summary of closely related literature to the civil unrest event prediction performed on various micro-blogging platforms such as Twitter and Tumblr. Naren et. al. [187] build a sequential probabilistic model based data analytics platform (called as EMBERS) that mine the data from various sources (tweets, news, blogs, web search, and Wikipedia) and generate warnings for civil unrest related events across 10 countries of Latin America. In extension to the study presented in Naren et. al. [187], Muthiah et. al. [188] describe planned protest model- one of the probabilistic models used in EMBERS for civil protest forecast. They propose a keyword based flagging and probabilistic model approach to learning about the date and location of the event. Zhao et. al. [190] use a dynamic query expansion technique and propose a local modularity spatial scan (LMSS) algorithm to identify general-interest and targeted domain events. Compton et. al. [191] propose a method for early detection of events (Latin America civil unrest, sports, public functions) based on the direct extraction of relevant and highly important tweets. In their proposed approach, they focus on four key phases: keyword based flagging to find relevant tweets, mention of future dates, event geocoding for identifying the location and logistic regression method to classify tweets for event detection. In extension to the study presented in Compton et. al. [191], Xu et. al [192] conduct the similar study on Tumblr website data for predicting civil disobedience related events. Hua et. al. [186] present an approach to collect tweets related to an event of interest instead of predicting events from tweets. They use a keyword based flagging approach and connect the dots between news reports and tweets based upon topic keyword matching. They divide their approach into two steps and first identify the key terms related to an event and the topic being discussed in the news. Using ranking algorithm, they find 200 top ranked topic keywords and in the second step of their approach they collect tweets published around the event date and containing those topic keywords.

Based on the literature survey, we find that existing studies use a variety of machine learning and data
mining techniques for predicting upcoming protest events. However, the prior research is conducted on the data or tweets posted during the event or after the event has happened increasing the possibility of bias in the dataset. Further, due to the high velocity and massive size of data uploaded on social media data, mining each and every post for building predictive model impedes the performance of the model. Motivated by the previous research and gaps, our aim is to address the challenge of noisy content present in the real time stream and building a model for investigating the potential of Twitter data as an open-source precursor for anticipating and predicting civil protests.

6.2 Research Contributions

In context to the existing work, the study presented in this chapter makes the following novel contributions:

1. We perform a content-based characterization and semantic enrichment on raw tweets to classify crowd-buzz & commentary and mobilization & planning microposts related to a given a protest or civil disobedience.[Characterization]

2. To the best of our knowledge, our work is the first case study on immigration dataset that presents a frequency based model for an early forecasting of events. We investigate the application of trend analysis (captured along the sliding window) for early detection of an event. [Predictive Modeling]

3. We conduct two case studies on the Twitter dataset by defining two events (Christmas- Island Hunger Strike- January 16, 2014 and Fast for Families hunger strike- December 12, 2013) and examine the effectiveness of our proposed solution approach. [Validation]

6.3 Proposed Solution Approach

In this Section, we present the general research framework and methodology of our proposed approach for an early prediction of civil unrest related events. We use Twitter microblogging website as a data source for conducting our experiments. Figure 6.2 illustrates the high level design and architecture of proposed method that primarily consists of three phases: Data Collection, Semantic Enrichment and Event Forecasting as labeled in the solution framework. We discuss all three phases in the following subsections.

6.3.1 Experimental Setup

In order to conduct our experiments, we download an open source Twitter dataset available on Southampton Web Observatory[^4]. Online Web Observatory is global data resource created for the Web Science research community [193]. The Web Observatory facilitates a distributed archive of data as well as the tools and mechanism to explore the evolution of web observatory. As discussed in Section 1.2.2, there are several web observatories created by large social media analytics groups from various universities. In the work presented in this Chapter, we conduct our experiments on the data acquired from one of such web observatories organized by Southampton University. We download Immigration Tweets[^5] dataset from the observatory consisting of approximately 2 millions of tweets spanning in a time duration of 5 months (from October 1, 2013 to February 28, 2014). As illustrated in Figure 6.2, this dataset is stored in MongoDB format. We query the web observatory data from our local terminal and convert the dataset from MongoDB to MySQL.

[^4]: https://web-001.ecs.soton.ac.uk/new/datasets
[^5]: https://web-001.ecs.soton.ac.uk/datasets/kxwiPvxKlxSEWkgsW
6.3 Proposed Solution Approach

Figure 6.2: A High-level Demonstration of Proposed Research Framework Primarily Consisting of 3 Phases: 1) Acquiring "Immigration Tweets" Data from Online Web Observatory, 2) Performing Semantic Enrichment on Tweets and Extracting Crowd-buzz & Commentary and Planning & Mobilization Tweets, and 3) Building A Frequency and Expression Correlation Based Model for Early Forecast of Civil Protests and Unrest Events.

Figure 6.3: Distribution of Locations of Users Discussing Event Related Tweets Collected During the Sliding Window Time Frame (7 Days)

by parsing the response given in JSON format. In the present study, we conduct experiments on English language tweets only. Therefore, we identify the language of the tweet posts using Java Language Detection Library\(^6\) and filter all records identified as non-English language tweets. The downloaded dataset is a collection of posts where each tweet consists of at least one of the following words: 'immigration', 'migration', 'immigrant' and 'migrant'. Since this dataset is restricted to 'immigration' related tweets, we identified the civil protest events related to immigration and happened during the period of data collection. We search popular news media and articles and find 2 such events. We provide a brief background about these events in the following subsections:

\(^6\)https://code.google.com/p/language-detection/
Table 6.1: A Sample of Related Keywords for Events 1 (‘Fast for Families’ at National Mall, USA) and 2 (‘Christmas Island Hunger Strike’ at Australian Detention Center)

<table>
<thead>
<tr>
<th>Event 1</th>
<th>Event 2</th>
</tr>
</thead>
</table>

6.3.1.1 E1- Fast for Families Hunger Strike, USA

A group of 4 advocates (Eliseo Medina, Dae Joong Yoon of NAKASEC, Lisa Sharon Harper of Sojourners, and Cristian Avila of Mi Familia Vota), all from different backgrounds set a tent in National Mall of USA on [December 12, 2013] and protested against new immigration reform bill. Later, in a month, 200 ordinary people joined them in the protest and approximately 10,000 people fasted across the country. This protest took place to show the urgency of new immigration reform for American families.

6.3.1.2 E2 Christmas island hunger Strike, Australia

The protest was initially sparked on [January 16, 2014] by the separation of some asylum seekers (almost 2000) from family members. In this protest, nine Iranian men stitch up their mouth with dental floss and threaten to sew up their eyes. This protest took place at Australian Detention Center.

In our proposed approach, we create a model for early detection of a civil unrest event, and hence we conduct experiments on the tweets posted before the events happened. We perform a manual analysis on Twitter and observe that the organizations planning for such protest or demonstrations mostly start publishing tweets a week before the protest date. Therefore, for each event, we use trend based sliding window of 7 days and extract tweets that are published during that time frame. For example, December, 5 to 11 for Event E1 and January, 9 to 15 for Event E2. In order to minimize the bias in our dataset, we extract the locations of users who posted the tweets present in our experimental dataset. Figures 6.3a and 6.3b shows the available locations of users who posted tweets related to Event 1 (December 5 to 11) and Event 2 (January 9 to 15) in 7 days time frame before the events happened. Figures 6.3a and 6.3b reveal that the dataset selected by sliding window is posted by the users belonging to different locations across the world and not specific to the locations of events- USA and Australia. The diversity of user location shows that our experimental dataset contains no bias. For Event E1 and E2, we were able to extract a total of 527 and 428 user locations respectively; among which 74 and 66 locations were discarded due to the presence of meaningless or non-informative locations. For example, 'YouTube', 'Facebook', 'Where hope floats' and 'Anywhere and Everywhere'.

6.3.2 Lead Indicator Classifiers

In this phase of proposed method, we perform various pre-processing and classification techniques on raw tweets and make them semantically rich to achieve better accuracy in event prediction. Based on our inspection of event related posts on social media platforms, we create a hypothesis that there are two types of posts that are shared before the event takes place.


6.3 Proposed Solution Approach

@Refugees how many asylum seekers are on hunger strike and other forms of self harm protest in Australian immigration detention centres?

http://t.co/ABrATCpsPN Several detainees at the Christmas Island immigration detention centre have sewn their lip...
http://t.co/lAnenwD7yW

@MinoWarrior @SherronShabazz We met with the organizers of the African migrant protests today in Tel Aviv

**Topic Location Temporal**

Figure 6.4: Examples of Civil Unrest Related Tweets Annotated with Temporal, Topic and Location Based Expressions

![Diagram of semantic relations between locations and topics](image)

Figure 6.5: An Example of Semantic Relations Between Locations and Topics in Event Related Tweets

1. **CrowdBuzz and Commentary (C&C):** In such posts people discussing about the event and spreading the word among other people who might participate in the event.

2. **Planning and Mobilization (P&M):** Group of people organizing the protest and sharing the posts for mobilizing and planning the event.

Due to the high velocity and massive size of data being posted on Twitter in real time, it is highly likely to have noisy and irrelevant posts that are not related to the particular event. Based on this hypothesis, we train a multi-class classifier that classifies tweets into above mentioned three categories. We use C&C and P&M tweets as two leading indicators (signals which help in early detection) for forecasting an event. These indicators filter relevant tweets to our problem of civil unrest event forecasting from all other tweets and hence reduces the traffic and extra computation of irrelevant tweets. We discuss the classification methods of these indicators in the following subsections:

### 6.3.2.1 CrowdBuzz and Commentary Tweets Classification

As discussed above, the crowdbuzz and commentary tweets are the posts that discuss the topic of the event. Therefore, for classifying such tweets, we identify the presence of pre-defined terms (lexicon-based approach) that are relevant to the event which is being monitored. We initially create a list of words that are important and commonly used in the events information. We further apply a bootstrapping method and extend the list by extracting hashtag and keywords from tweets posted in seven days sliding window time frame. We extract event related information by mining "Google News" media websites and articles related to the event and further enrich our list by using TF-IDF (Term Frequency-Inverse Document Frequency) based approach. Table 6.1 shows the list of related keywords for event 1 (Fast For Families- USA) and event 2 (Christmas Island Hunger Strike, Australia). If a tweet present in the selected time frame of the event contains any of the event-related keywords, then we classify the tweet as crowdbuzz and commentary tweet.

### 6.3.2.2 Planning and Mobilization Tweets Classification

As discussed above, the planning and mobilization tweets are used to organize and plan an event and hence contains the information about the location and time information about where and when the event is going to take place. Therefore, for classifying P&M tweets, we propose to extract the spatiotemporal
feature from tweets. We also observe that P&M tweets contain reply or direct mention to other users [@] character followed by a Twitter username due to the exchange of messages between users for coordinating their activities. We extract the person, location and temporal features from the tweets using an ensemble based learning on a combination of Named Entity Recognizers. We use three open source NER APIs for the purpose of features extraction: Java NER \(^9\), SUTime \(^10\) and TextRazor \(^11\). We observe that due to the presence of free-form text and user-generated data, there is no defined structure of the tweets and hence contain inconsistency. For example, while some tweets mention the date of the event, some tweets contain the day information while some tweets mention future terms (tomorrow or day after tomorrow) for temporal information. Therefore, to maintain the consistency in temporal expressions, we convert the extracted expressions into the day of the week based upon the timestamp of the original tweet. For example- if a tweet contains the entity ‘tomorrow’ and is posted on ‘Wednesday, January 15, 2014’, then we convert it into ‘Thursday’. Another feature that we come across to is that the tweets posted for planning and mobilizing consist of phrases like join us, spread the word and future tense related words. We annotate 500 P&M tweets and train a machine learning classifier for filtering the P&M tweets. We discard the remaining (not identified as C&C or p&M) as irrelevant tweets. Figure 6.4 shows concrete examples of location, temporal and topic expressions in tweets related to civil protest events.

Figure 6.6 illustrates the trend of tweets being posted in every hour for 7 days sliding window time frame. Figure 6.6 reveals that in comparison to the total number of tweets posted in a time frame, only a very small fraction of those tweets are C&C and P&M which increase the significance of filtering these posts. Figure 6.6 shows that the number of C&C tweets has similar pattern for five days before the event happened and sudden peak for last two days, unlike P&M tweets which are higher in initial days of planning. We also observe that the event happened in Australia. However, the maximum number of tweets during this time frame are posted from various locations of USA. This trend shows that it is not a good idea to predict the location of an event based upon the user profile locations with the maximum number of tweets posted.

\(^9\)http://nlp.stanford.edu/software/CRF-NER.shtml  
\(^10\)http://nlp.stanford.edu/software/sutime.shtml  
\(^11\)https://www.textrazor.com/docs/java
6.3 Event Forecasting Model

We define the problem of civil unrest event forecasting as applying OSSMInt for extracting three primary insights (TTL) from the event-related tweets: Temporal expression $T_i$- time or date of the event, Spatial expression $l_o$- location of the event, and Topic expression $T_o$- root cause or objective of the protest. As shown in Figure 6.2, in order to build an event forecasting model, we take an input of crowdbuzz-commentary and planning-mobilization tweets that are enriched with TTL information extracted using ensemble based named entity recognizers and TF-IDF approaches. In our proposed approach, we implement an adaptive version of the algorithm proposed in Budak et. al. [194]. In addition to the spatiotemporal features proposed in Budak et. al. [194], we add topic as another discriminatory feature for event prediction. Figure 6.7 shows the name of all topics identified from the tweets present in our experimental dataset of Event 2 (Christmas Island Hunger Strike, Australia) sampled for a time frame of 7 days of sliding window.

A tweet might contain more than one named entity expressions. Therefore, we define a bipartite graph arrangement of all location, time and topic expressions. A bipartite graph is an arrangement of vertices ($V$) and edges ($E$) where $V$ is split into two disjoint sets of vertices $V1$ and $V2$. An edge in a bipartite graph can be represented as $e \in E = \{v_1, v_2\}$ where $v_1 \in V1$ and $v_2 \in V2$. For each topic $To_i \in To$, location $Lo_j \in Lo$ and temporal expression $Ti_k \in Ti$, we create distinct set of all possible combinations of expressions pairs: location-topic $<Lo_j, To_i>$, topic-time $<To_i, Ti_k>$ and location-time $<Lo_j, Ti_k>$. We use these named entities pairs to compute the correlation among a location, time and topic. To keep only significant results, we select only those expressions that are at least $\theta$ frequent (occurs at least in the $\theta$ number of tweets) where $\theta$ varies for all 3 entities. We define $\theta$ for each entity as the average number of named entity expressions present in the dataset (only C&C and P&M tweets). Figure 6.5 reveals the significance of selecting $\theta$ frequent entities for event prediction. The network graph demonstrated in Figure 6.5 shows a sample of topic entities collected from the C&C and P&M tweets for Event E2, posted on the first day of the sliding window. Figure 6.5 reveals that the terms ‘protest’, ‘refugee’, ‘African people’, ‘immigration’, ‘Tel Aviv’ are the terms which are significantly correlated and highly frequent in tweets in comparison to ‘racism’ or ‘prostitution’. To find these significant pairs, we extract all expressions of $x$ entity that are at least $\theta$ frequent in the dataset $F(x) > \theta$

and further extract all those expressions of paired entity $y$ that are $\psi$ frequent for at least one $x$ and vice-versa. $F(x, y) > [\psi F(x)] \quad F(y, x) > [\varphi F(y)]$

Here, $F$ is the frequency of an entity expression in the dataset. We select the pairs of named entities which satisfy the above conditions, and it reduces the number of pairs for further examination. We compute the
frequency of each pair in a bipartite manner and only look for the pairs that are highly correlated and has no decrement in the frequency in 7 consecutive days of the sliding window. Since these named entities are categorical attributes, we use Chi-Squared distribution to find the correlation between two entity expressions. We define a pair of expressions to be significantly correlated if their p-value < 0.05 for their respective $\chi^2$ and degree of freedom.

6.4 Experimental Results

6.4.1 Crowd-buzz and Mobilization Tweets Classification

To evaluate the performance of our proposed approach, we use standard measures of information retrieval and machine learning. We asked 50 graduate students to manually annotate each tweet of the sliding window data. We provided them specific guidelines for annotating crowd-buzz, planning & mobilizing for the general event and domain specific event (Fast for Families Protest- December 12, 2013 and Christmas Island hunger strike- January 16, 2014). Based on their decisions we validate our results and compute accuracy of semantic enrichment phase using precision, recall, and F1-score. Tables 6.2 shows the standard confusion matrix for both C&C and P&M classifiers for Events E1 and E2. NA denotes the number of tweets classified as irrelevant or unknown. Table 6.2a reveals that out of 1,846 tweets (2.3% of sampled tweets for Event 1- 79,431) annotated as C&C, we were able to predict 1,322 tweets correctly while 77,223 tweets as irrelevant with a total misclassification of 1.1%. Similarly, Table 6.2b reveals that out of 828 tweets (less than 1% of sampled tweets for Event 2- 82,962) annotated as crowdbuzz, we were able to predict 719 tweets correctly with a total misclassification of less than 0.5%. Tables 6.2c and 6.2d shows the confusion matrix for planning and mobilization tweets identification. As demonstrated in Tables 6.2c and 6.2d, since the presence of spatiotemporal features is not specific to civil unrest events; we not only identify the P&M tweets that belong to protest-related events but also any general events. For example, music concerts, seminars, conferences, official government event. Tables 6.2e and 6.2d reveal that for both events E1 (274 P&M tweets) and E2 (142 P&M tweets), we were able to correctly classify 140 and 127 tweets respectively. Table 6.3 shows the accuracy results of crowdbuzz & commentary and planning & mobilization classifiers.
Table 6.3: Accuracy Results for Semantic Enrichment Classifiers (Crowd-Buzz and Mobilization Tweets)

(a) Fast For Families

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>C&amp;C</td>
<td>0.78</td>
<td>0.71</td>
<td>0.75</td>
</tr>
<tr>
<td>P&amp;M</td>
<td>0.51</td>
<td>0.93</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>0.91</td>
<td>0.86</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>

(b) Christmas Island Hunger Strike

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>C&amp;C</td>
<td>0.84</td>
<td>0.86</td>
<td>0.85</td>
</tr>
<tr>
<td>P&amp;M</td>
<td>0.96</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Figure 6.8: Distribution of $\chi^2$ and p-value for Frequent Pairs of Locations and Topics for 3 consecutive Days in Sliding Window- Christmas Island Hunger Strike (CIHS) and Fast For Families Protest (F3P)

for events E1 and E2. Table 6.3 reveals that the precision, recall, and f1-score for both the classifiers are reasonably high and we are able to classify crowdbuzz and mobilization tweets for events E1 and E2 with the f1-score of 0.75 and 0.85 respectively.

6.4.2 Chi-Squared Distribution

6.4.2.1 Event 1: Fast For Families

For "fast for families protest in national mall" event related tweets, we find 8 unique and non-overlapping topics, 3 unique locations and 3 unique temporal expressions that have a relatively high non-decreasing frequency in tweets posted during sliding window time frame (7 days). In order to identify the significant pairs of entities, we compute $\chi^2$ and p-value distribution for each frequent pair. Chi-squared distribution ($\chi^2$) is a statistical test applied to the categorical data to compute the frequency distribution of certain events observed in a sample data [195]. Here, the events are defined as the entities (spatiotemporal and topic entities). The p-value is computed as an evidence of the statistical model and defined as the probability of obtaining the results same or more extreme than the real observed results [195]. We compute $\chi^2$ distribution of all frequent pairs and discard the ones that are less frequent than $\psi$ threshold value. We further normalized all at least $\psi$ frequent pairs of topic-time-location entity expressions between 0 and 1 (similar to the first event). Based upon the distribution, we observe that the pairs containing location "Washington (0.034)" has
very less p-value in comparison to "Europe (0.23)" and "California (0.014)". Figure 6.8a shows the variance between p-value of location "Washington D.C." and all frequent topics for three consecutive days (December 2 to 4).

### 6.4.2.2 Event 2: Christmas Island hunger strike

For "Christmas Island hunger strike" event, we find 10 unique topics, 6 unique locations and 5 unique temporal expressions in our dataset. Based upon the frequency model (discussed in Section 6.3.3) we find only 3 unique topics ('right of asylum', 'migrant worker' and 'refugee'), 4 unique locations ('Australia' and 'US') and 2 unique temporal expressions ('Wednesday' and 'Thursday') with no decreasing frequently in sliding window. We further compute p-value of each pair based upon the \( \chi^2 \) and degree of freedom and observe that the pairs containing 'U.S.' have p-value greater than 0.05 while Australia has p-value to be very small (approx 0.000002). Figure 6.8 illustrates the variance between p-value of 'Australia' and 'US' for all topics for three days in sliding window (p-values are normalized between 0 and 1). Similarly, for (time, location) and (time, topic) pairs we keep highly correlated terms with significant p-values and are able to predict events with a high confidence score.

### 6.5 Conclusions & Future Work

In this work, we present an approach for early detection and prediction of civil unrest related events. We conduct experiments on real-world dataset (open source dataset downloaded from Southampton Web Observatory) consisting of tweets on immigration and migration. We use named entity recognition and term-frequency based approaches to capture various discriminatory features from raw tweets such as time, topic and location of the event. We perform single and multi-class classifications to filter event related tweets (crowd buzz and mobilization). Experimental results reveal the high accuracy of classifiers with an f1-score (0.75 and 0.85 for "Fast for Families protest" and "Christmas Island hunger strike" events respectively). We develop a frequency based model on these semantically rich tweets and find those pairs of location, topics, and temporal based named entities/expressions that are significantly correlated. We further compute the \( \chi^2 \) and p-value distribution of these entities and their correlation over a time frame. From this distribution, we conclude that by detecting trend analysis of these entities in the tweets posted during a sliding window time frame, we can predict civil unrest related events with a high confidence score. We also conclude that early identification crowdbuzz and mobilization tweets are value added in event forecasting.

Future work includes the extension of proposed method for detecting protest related events in real time and events with an overlapping sliding window (multiple events occurring on the same day or in the same week). We plan to investigate the application of Ensemble Learning approach for event forecasting and semantic enrichment. Furthermore, future work includes the timeline-based visualization of tweets to provide end-users an updated and rich information (highlights or summary) about events in real time. Our future work also includes the early prediction of a protest or civil disobedience based on the analysis of a chain of similar incidents already happened in past or currently happening and tends to happen at another location or time for the similar root causes.