# Spatio-Temporal Mining of Social Unrest Events

Bill Mutabazi*, Minal Khatri*, Nancy Pham*, & Venkata Krishna Mohan Sunkara*
*Computer Science and Engineering Department, University of Nebraska - Lincoln

***Abstract*-- Social unrest events, which are common happenings in both democracies and authoritarian regimes, are complex events that have a high risk factor associated with them and have impact on huge masses of people. In this project , we explore the different causes for these unrest events and the similarities between different types of social unrest events by applying data mining techniques like association-rule mining and clustering. We combine three different set of features: (1) Spatial -temporal (2) Socioeconomic (3) Infrastructure that can influence a social unrest event. In addition, we explored association rules which provide meaningful insight into the different factors associated with an unrest event using apriori algorithm and visualize the events in an effective way. We also use this huge heterogeneous dataset to cluster events which are closely related to each other in either space or time or socio-economically and visualize them to observe the clusters and outliers. Extensive experiments with data from seven states in India demonstrate the effectiveness of this project. We also perform text-mining on newspaper articles from Times of India and obtain the event locations. This project defines how different events that are associated in spatial and temporal domain could be mined by the use of different data mining techniques which could help in prediction of events and better security assurance against these social unrest events.**

## 1. INTRODUCTION

Social unrest events (protests, strikes, demonstrations, and occupation) are common happenings in both democracies and authoritarian regimes [6]. Most social unrest events are 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 India as an example: The Jat Reservation Agitation was a series of protests in February 2016 by Jat people of North India, especially those in the state of Haryana, which "paralysed the State for 10 days." Anticipating these latent instabilities before they occur and applying preventive strategies to avoid them have important ramifications, such as prioritizing citizen grievances for the decision makers, issuance of travel warnings for the tourism industry, and insight into how citizens express themselves for the social scientist, which has motivated many social and data science researchers to focus on revealing the patterns contained in these events and further the prediction of future latent social unrest.

Spatial data mining, i.e mining data from large amounts of spatial data, is a highly demanding field because huge amounts of spatial data have been collected in various applications. A spatial contains objects which are characterized by a spatial location and/or extension as well as by several non-spatial attributes. [1]

By the complexity of spatial data type, our project will explore and apply various data mining techniques on census data and geographical data for a large spatio-temporal data set of seven different states in India. By using the Geographic Information System (GIS) as a framework for gathering, manipulating, and analyzing data, users can query spatial data and perform simple analytical tasks using programs or queries. However, the GIS does not perform complex data analysis. In our project, we will implement various data mining tasks such as clustering, text mining, and trend determination for existing data of social unrest events in India.

In reference to social unrest events, the following points are problematic and need analysis: (1) It is unclear how different causes and combination of causes can affect the type of social unrest event. (2) It is unclear how different events are associated in terms of space and time domain. (3) It would be interesting to find out similarities between different unrest events as that can help to decide on common preventive measures. Analysing these problems and identifying the relevant factors can be helpful to mitigate the impact of these events.

**Figure 01: Map of India.**

By applying various data mining techniques and algorithms, we will analyse the data basing on different attributes to identify the trend and factors in which most of the unrest events have happened in the previous years. Our work could serve as a planning tool to forecast future events and crime prevention in the predetermined regions ( seven states of India).

Figure 01 shows the seven states of India that were considered for this project. The green-colored states are those that were considered; whereas the red-colored states were not. For the text mining objective, all states were considered.

## 2. OBJECTIVES

Our four objectives for this project are as follows:

- **Data acquisition and cleaning** of various data obtained from GDELT, Census and Shape files.
- **Text Mining** on web articles to get the social unrest event location along with the context of the event.
- **Association Mining** to get association rules among different attributes and the type of social unrest event
- **Clustering** using a conceptual distance function to obtain similarities among data along with outliers in the data.

For data acquisition and cleaning, we obtained shapefiles of infrastructure points, census data, GDELT data, and combined these files. For text mining, we scraped web news articles for social unrest events based on keywords. From this, obtained the spatio-temporal coordinates and mapped the events. For association analysis, we considered socio-economic, spatio-temporal, and infrastructure attributes. For clustering analysis, we used a conceptual distance function to obtain the similarity between events.

## 3. WHAT IS CIVIL UNREST?

Event analysis of the form considered here is an established concept in social science research [7]. Civil unrest is a large concept by which people express their protest against things that affect their lives. If the action is directed against private actors, there is normally a connection to government policy or behavior, e.g., a labor strike against a private company can disrupt the rhythm of everyday life for the rest of society, turn violent or lead to a series of disruptive strikes which require government involvement, and thus responsibility in the eyes of citizens. Civil unrest does not include acts by criminals for purely private gain. While authoritarian governments may outlaw civil protest and thus criminalize the participants, social scientists would distinguish illegal political protests from illegal criminal activities. Gang members stopping public buses to extort payoffs from bus owners would not be a civil unrest event, though people protesting afterward against the government's inability to control such gangs would be considered civil unrest. Regardless of a country's level of openness to citizens' expression, civil unrest may occur in carefully planned and orchestrated forms or as spontaneous responses to external events.

This expansive definition of civil unrest means that one can find it everywhere, including European protests against austerity or marches against an oil pipeline from Canada across the US to the Gulf of Mexico. India, nevertheless, offers some special characteristics that make it an excellent region for study in our project. The region experiences a plethora of civil unrest events every day, is well covered by international and national news media (facilitating the task of generating ground truth), is the object of detailed empirical research and polling (permitting the description of the social, political and economic context within which civil unrest occurs) and has a significant and growing number of social network users (thus supporting the use of modern data mining algorithms). [8]

## 4. RELATED WORK

Event extraction, where structure description of events are codified from text (Newspaper articles). Baeza-Yates [9] provides a ranking of future events retrieved from news. Spatial and temporal distributions of civil unrest over 170 countries were studied in [10]. Event prediction has been explored in a variety of applications, including elections [11], disease outbreaks [12], stock market movements [13], social unrest event prediction [14, 15]. 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 do not 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 are announced in advance [16]. Time series based mining uses temporal correlation of relevant features such as tweet volume by adopting appropriate approaches. Radinsky and Horvitz [17] 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. [19]

The GDELT Project [18] is a real time network diagram and database of global human society for open research which monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, counts, themes, sources, emotions, counts, quotes, and events driving our global society every second of every day, creating a free open platform for computing on the entire world. Each day the GDELT Project monitors the news media across nearly every corner of the world and compiles a list of over 300 categories of "events" from riots and protests to peace appeals and diplomatic exchanges, recording the details of the event, including its georeferenced location, into a master "event database" of more than a quarter-billion events, dating back to 1979 and updated each morning around 4AM EST.

In particular, from 19 February, 2015, GDELT 2.0 has been online which updates every 15 minutes accessing the world's breaking events and reaction in near-real time. In GDELT event data table, each record has 58 fields (61 fields in GDELT 2.0), capturing information pertaining to a specific event in CAMEO format [35]. In this paper, we use the following nine fields from a record: SQL DATE, Month Year, EventRootCode, Goldstein Scale, NumMentions, AvgTone, ActionGeo_CountryCode, ActionGeo_Lat, and ActionGeo_Long. SQL DATE and Month Year

are the date the event took place in YYYYMMDD format and YYYYMM format, respectively. EventRootCode defines the root-level category the event code falls under. For example, code 1452 (engaging in violent protest for policy change) has a root code of 14 (PROTEST). This makes it possible to aggregate events at various resolutions of specificity. Goldstein Scale is a numeric score from −10 to +10, capturing the theoretical potential impact that type of event will have on the stability of a country. NumMentions is the total number of mentions of this event across all source documents, which can be used as a method of assessing the importance of an event: the more the discussion of that event is, the more likely it is to be significant. Avg Tone is the average tone of all documents containing one or more mentions of this event. The score ranges from −100 (extremely negative) to +100 (extremely positive). ActionGeo_CountryCode is the location of the event, which is a 2-character FIPS 10-4 country code for the location. ActionGeo_Lat and ActionGeo_Long are the centroid latitude and centroid longitude of the landmark for mapping. The dataset is also available on Google Cloud Platform (https://cloud.google.com/) and can be accessed using Google BigQuery. In this paper, we export the following GDELT event data for the experiments from the Google BigQuery (https://bigquery.cloud.google.com/table/gdelt-bq: full.events?pli=1) web service

In this project, we explore web based articles, particularly for 7 states of India. With the recent increase in the use of social-media, "big data" (e.g., via channels like Twitter, Facebook, Youtube) has given a new window into studying events across the globe. As a result, social media data has grown enormously over the past few years. Through the use of data mining techniques, different researches have been able to aggregate public data to capture triggers underlying events, detect on-going trends, and forecast future happenings. Concomitantly, there has been a rapid development of new computational methods for spatio-temporal mining of social media datasets [2].

## 5. PROBLEM DEFINITION

Social unrest events are complex events that have a high risk factor associated with them and have impacts on huge masses. Risks can generally be understood as the potential for experiencing harm. More specifically, it denotes the likelihood of a scenario leading to adverse effects caused by an activity, event, or technology [4]. In reference to social unrest events, following points are problematic and need analysis: (1) It is unclear how different causes and combination of causes can affect the type of social unrest event. (2) It is unclear how different events are associated in terms of space and time domain. (3) It would be interesting to find out similarity between different unrest events as that can help to decide on common preventive measures.

Analyzing these problems and identifying the relevant factors can be helpful to mitigate the impact of these events. Therefore, in this work , we obtain the social unrest event dataset for 7 different states of India and perform association mining to analyze relation among attributes or combination of attributes with the type of social unrest event and clustering to identify the similarities between different types of event. Based on the results of clustering and association mining, domain experts can decide on preventive measure and make policies to mitigate the impact of these events.

## 6. DATASET

Data is collected from various sources and filtered for the 7 states which we are focusing and then pre-processed and combined together to a final dataset. The final dataset consists of 90 numerical attributes and 20 categorical attributes. Of the 20 categorical attributes, 15 were related to a location. So, the final dataset which is being used for association analysis and clustering consists of a total of 84 attributes out of which 83 are numerical attributes and only one (protest_category) is the categorical attribute. For the text-mining objective we have collected data from the archives of a website for the year 2017. It consists of approximately 8000 unrest articles over a period of 4 months..

To obtain the final dataset, We combine the data pre-processed from Indian census data, the Global database of society data (GDELT), which is an open website that records and monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, themes, sources, emotions, counts, quotes, images and events driving our global society dynamically for every second every day, creating. In addition

since this data contained some noise and had many unnecessary data regarding our objectives to fulfill our purpose we added from the shapefiles using Hot Export Tool for 7 different states in India, including Bihar, Karnataka, Maharashtra, Kerala, Sikkim, Goa, and Gujarat which are the states that our project focuses on.

Some of the infrastructure points considered have the following infrastructure attributes: schools, colleges, universities, tourism spots, restaurants, police stations, fire stations, banks, government offices, military, hospitals, parks, residential areas, gas stations, places of worship, libraries, historic centers, stadiums, sport centers, airports, train stations, and bus stations. These infrastructure point were chosen from a set of many other infrastructure point based on how important they could be related to the social unrest event.

Preprocessing steps include cleaning all of the data and combining all of the separate shapefiles. In addition, we downloaded the amenities data, and we combine this with existing census data.

For our GDELT event data, we have 39,448 records instances and 83 attributes. The main categories of attributes are spatial and temporal attributes, socio-economic attributes and Infrastructure attributes.

## 7. APPROACH

### 7.1 Data Collection and Preprocessing Data.

In this section, we discuss the data collection and data preprocessing steps in detail. The most challenging part is to combine data from different sources.

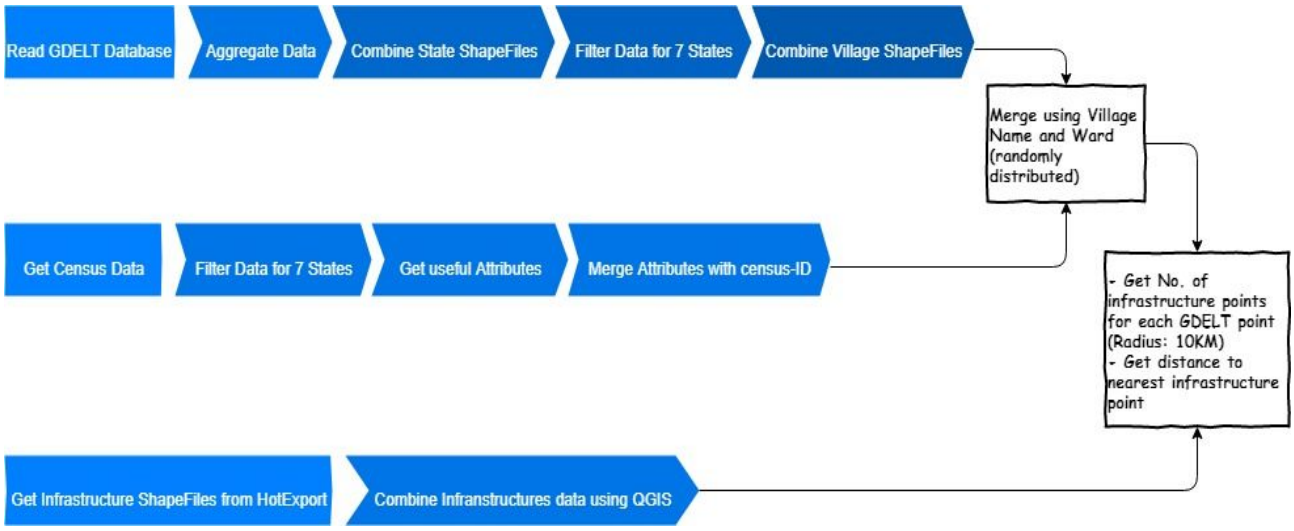



**Figure 02: Data collection and pre-processing**

The data is collected from three different sources which includes:

- Global Database of Events Language and Tone (GDELT) database
- Census database
- Open Street Maps

GDELT database as a dynamic open source database for social unrest events all across the globe. Census database consist of census attributes like :population data , amenities data etc. Most of the countries have restrictions on the access to geographical data but Open Street Maps is a open source database for downloading raw geographical data.

The data collected from various sources is cleaned using preprocessing techniques which includes:

- Python, Java and R scripts
- QGIS application

Python scripts are used for filtering and aggregating data. Java scripts are used for accessing the database. For combining shapefiles, geo-spatial packages in R programming language is used. QGIS application provides a platform for viewing, editing and analyzing geospatial data .

Figure 02 presents the systematic view of the data collection and preprocessing steps.

**Stage-1**: A java script was written to download data from GDELT database. The database includes location attributes like: latitude and Longitude of event., temporal attributes like: day, month and year of event and 7 social unrest event categories: Appeal, Assault, Coerce, Demand, Fight, Threaten, UMV(Ultra Mass Violent) and other attributes like: Country name, State name, City name in which the event took place.. The State Name attribute has a lot of inconsistent and incorrect data therefore, we discard the attribute. In the Next step, we aggregate data for enhancing the semantics, by adding a count attribute which maintain the number of same type of events at same location on the same day. Further, we combine the state shape files obtained from census website. Then, we filter the data for 7 states. After that we combine the village shapefiles. Finally, we obtain the database with spatial temporal attributes. In the next, stage we discuss, about the census database.

**Stage-2**: A java script was written to download data from Census website which consist of following attributes:

1. Population data
2. Housing Attributes
3. Amenities (Village)
4. Amenities (Urban)

In the next step, a python script was written to filter data for 7 states. There were many attributes ,which were not relevant for the domain like : type of houses, number of toilets so, we filter these attribute and get the useful attributes which were relevant for the domain. Finally, we get the socio-economic attributes.
Now, at this stage we merged the GDELT database and Census data using 'village_name' attribute as the similarity measure and for wards , we randomly distribute the events across wards based on city _name in GDELT.

**Stage-3**: Using the Quantum Geographic Information System (QGIS), an open source cross-platform that supports viewing, editing, and analysis of geospatial data, we combined data for the identified states of india downloaded from Openstreetmap (OSM) and loaded those files into QGIS to get distance of the nearest infrastructure point data for every social unrest event. Finally, we merged data loaded in QGIS and set attribute of the nearest radius from one event of a given protest category to another to 10 Kilometers (Km) and the resulting dataset of 83 attributes and 39,448 instances was our final Dataset used for our association, clustering and text mining.

**7.2 Text Mining**.

- **Web Scraping** to collect all web articles for 2017 from Times of India newspaper
- **Obtain social unrest articles** by filtering all articles
- **Geocoding** locations of social unrest events

The newspaper articles were collected from Times of India, a major newspaper website in India, using a Java package. The Java file accesses the archives of the news articles. It saves each newspaper article into a separate text file. The web articles were collected for the year of 2017, resulting in approximately 220,000 articles.

This data is then processed using natural language processing techniques and the Natural Language Toolkit (NLTK) package in python. A program was written, using built-in functions from NLTK, to pre-process the data and filter out all articles unrelated to social unrest events. Specifically, the functions corrected typos in the web articles. In addition, different tenses of a verb were considered. For example, "murdered," "murdering," and "murder," were all considered to be one word. This step was important for filtering articles for social unrest events. The filtering was completed using a predefined vocabulary list of approximately 60 words associated with unrest events, such as riot, demonstration, protest, fight, government, etc. A program was written to count up all instances of these words in each article. This count was then summed up to determine how many social unrest words were contained per article. If an article had at least 5 of the words from the predefined vocabulary, it was categorized as a social unrest article. The sum of 5 was chosen after some trial and error. When this was tested with a sum of 3, there were still articles leftover related to other subjects, such as movie articles. The sum was then increased to 5 to filter out such articles and only keep social unrest articles. This was applied to the months of January, February, and March, and resulted in approximately 12,000 social unrest articles.

After obtaining the filtered social unrest articles, the Geograpy and Geopy packages in python were applied. These packages obtain the nouns from the articles that correspond to a place. It then outputs a string of nouns in each article. This string of nouns is given to Google Maps API. Google Maps API takes the string of nouns and outputs a latitude and longitude for that location. This latitude and longitude is then mapped using the Folium package in python.

**7.3 Association Mining**.

- **Dimensionality Reduction** to reduce the number of attributes to 10 from 83.
- **Normalization** with respect to population
- **Discretization** using equal frequency binning (3 bins)
- **Association analysis** using apriori algorithm.

The data obtained from the pre-processing step consists of a total of 83 attributes and the apriori algorithm cannot handle data with more number of attributes. So, a dimensionality reduction technique is applied to condense the data in to 10 attributes using Principal Component Analysis (PCA) or using the formula described below where d1, d2, ----, dn are the attribute values. The attribute values provided to the formula are converted to a single value which is considered as a reduced representation of all attributes.

$$1- \frac{\sqrt{(1-d1)2+(1-d2)2+(1-d3)2+ \ldots.+(1-dn)^2}}{\sqrt{n}}$$

The reduced dataset now consists of just 10 attributes which are described in table 01. These attributes are then normalized with respect to population and scaled with in the range of 0-1. As association analysis requires all attributes to be categorical, they are discretized using equal frequency binning technique into 3 bins. Now apriori algorithm is applied on the discretized data with a support of 0.01 and confidence of 0.01.

PCA is also applied on a subset of features to obtain a single attribute and this technique is applied on different subsets of features to obtain a total of 10 attributes. The same process as above is followed to obtain the association rules.

**Table 01: Attributes in Association Analysis**

| Attribute Name | Description |
|---|---|
| TRU | Urban or Rural |
| Count | Number of events at that location on the same day. |
| protest_category | Protest category as obtained from the GDELT database. |
| Social status | Social status of a particular region (Village/Ward) obtained from various infrastructure counts. |
| population | All attributes related to population of a specific region (Village/Ward) |
| infra_area | The number of infrastructure points in a region (Village/Ward) |
| social_housing | The housing attributes of a region obtained from the census data. |
| Distance | The distance from a region (Village/Ward) to the district and sub-district headquarters. |
| infra_dist | The shortest distance to 8 different infrastructure attributes within a 10km radius |
| infra_num | The number of infrastructure points within a 10km radius. |

After obtaining the association rules, the rules are filtered such that either the event count or protest category is present on the right hand side of the rule. Then the instances supporting the first 3 rules are obtained which are then displayed on a map using the geographical coordinates of those instances.

**7.4 Clustering**.

- **Spatial distance** is calculated using the haversine formula [2].
- **Temporal distance** is calculated by obtaining the number of days in between two event dates.
- **Infrastructure distance** is obtained by considering the absolute difference between corresponding attributes (all remaining attributes from table 01) of two events.

The distance function which is used for clustering similar events is obtained by taking the average of spatial, temporal and infrastructure distances as shown in figure 03.

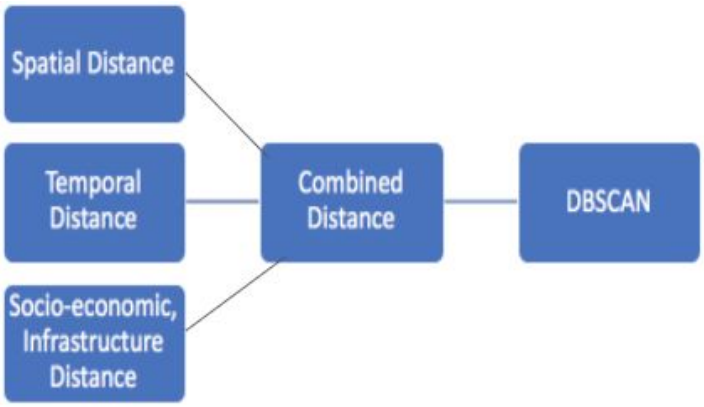



**Figure 03: Clustering of similar events.**

The combined distance function is a representation of all original 83 attributes. This combined distance provides a conceptual meaning between two events. Hence, if two events are close to each other based on the combined distance function, then they are probably similar to each other. Based on this intuition, DBSCAN algorithm is applied on the reduced dataset (10 attributes) to cluster events which are similar to each other.
Clustering is done using just the spatial distance, spatio-temporal distance and also using the combined distance. The instances which are grouped together in a cluster are obtained and displayed on a map using the geographical coordinates of those instances.

## 8. RESULTS

We rely mostly on python and R programming languages to implement these objectives. We use several inbuilt packages of python and R for data analysis and cleaning. We use the arules package in R and scikit learn package in python for association analysis and clustering respectively. We also utilized the jsoup package in java for accessing the news articles and the NLTK package for text mining. For spatial analysis and combining different shapefiles we utilized the QGIS software.

**8.1 Text Mining Results**.
After web-scraping and filtering out all articles unrelated to social unrest, the locations of the social unrest events were obtained and geocoded. Figure 04 shows the geocoded social unrest events for January, February, and March 2017. The results are as expected: a majority of events happen in bigger cities with higher population, and less social unrest events happen in areas with less people.

We manually observed some of the articles and found that almost 80% of the articles are somehow related to unrest events are also being geocoded correctly. For example, we observed an event related to student protest at a university in Hyderabad and when we observed the geocoded position we found that the article is correctly geocoded.

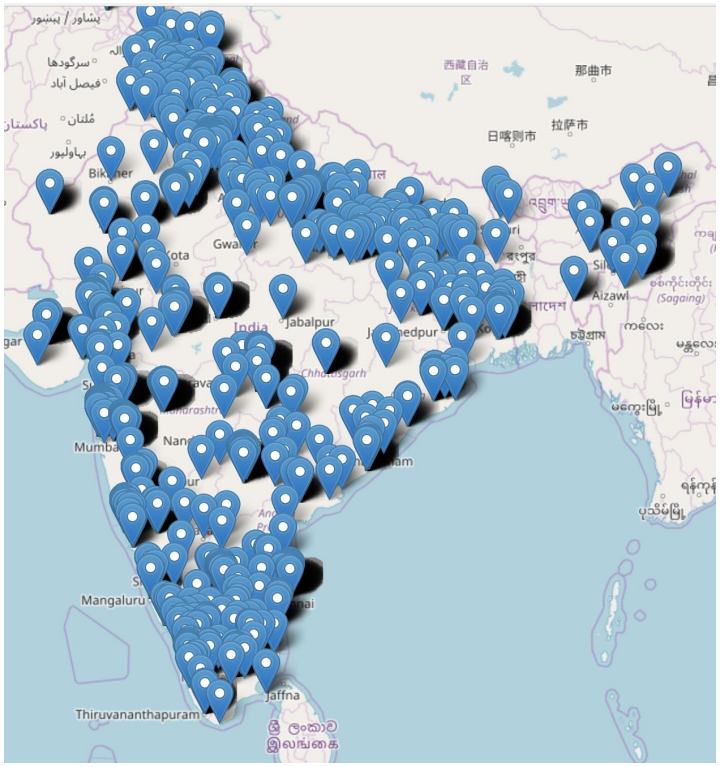



**Figure 04: Geocoded social unrest events for January - March 2017**

**8.2 Association Analysis Results.**
Several rules were obtained by varying the support and confidence values as shown in table 02 and table 03 . However, all the rules generated are not interesting and we require only the rules containing either count or protest category on right hand side of the rule. When tried with equal width binning, 0 rules are generated with either count or protest category on the right hand side of the rule as all the instances belong to a single bin. So, we tried using the equal frequency binning to obtain some meaningful rules.

**Table 02: Number of rules generated using the dimensionality reduction formula**

| SUPPORT | CONFIDENCE | #RULES |
|---|---|---|
| 0.5 | 0.8 | 0 |
| 0.5 | 0.5 | 1 |
| 0.3 | 0.1 | 35 |
| 0.01 | 0.01 | 88817 |

**Table 03: Number of rules generated using PCA**

| SUPPORT | CONFIDENCE | #RULES |
|---|---|---|
| 0.5 | 0.8 | 0 |
| 0.5 | 0.5 | 1 |
| 0.3 | 0.1 | 35 |
| 0.01 | 0.01 | 89757 |

Below we describe one of the interesting rule generated having count on the right hand side.

{protest.category=Fight, Social.Status=poor, population=more, social_housing=better, infra_num=good, infra_dist=near} → {count=less} 2.73 584

The rule states that if the protest category is fight and the social status as determined by some census attributes is poor and the population attributes like literacy rate, working population are more, and housing, amenities in that village/ward is better, and the infrastructure with in a 10km radius is good and are near to the event then the count is less. This rule has a lift measure of 2.73 and 584 instances support this particular rule. Figure 05 shows a map of instances which supported this particular rule.

Figure 05 shows only 3 major locations as most of the events which support this rule occurred at those locations. The locations represented on the map are all urban regions which have better infrastructure and more population. However, one might expect to observe more number of events in those locations but in contrast the rule states that the event count at those locations is less.

Below we describe another rule which is generated having protest category on the right hand side.

{Social.Status.=good, pop=less, distance=accessible, infra_dist=far} → {protest.category=Fight} 1.38 440

The rule states that if the social status as determined by the census attributes is good and the population attributes such as literacy rate, non-working population are less and the distance to district and sub-district headquarters is accessible and the nearest distance to infrastructure points is far, then the protest category is more likely to be a fight. This rule has a lift measure of 1.38 and 440 instances support this particular rule.

Figure 06 shows the instances which support this rule and we can observe that it isn't concentrated around a particular region but rather dispersed as the attributes as mentioned above are not common and can occur at any of the villages or wards. Hence the instances that support this rule are pretty diverse.

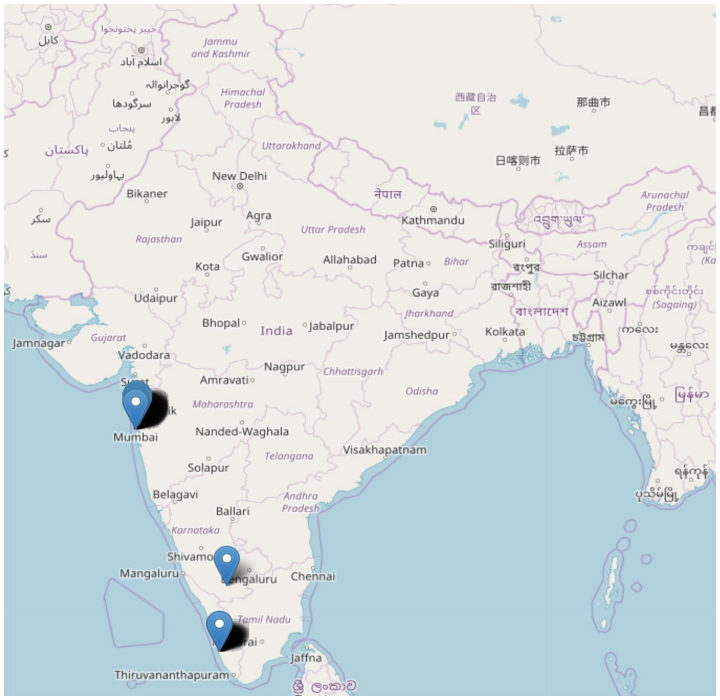



**Figure 05: Instances which support the rule related to event count.**

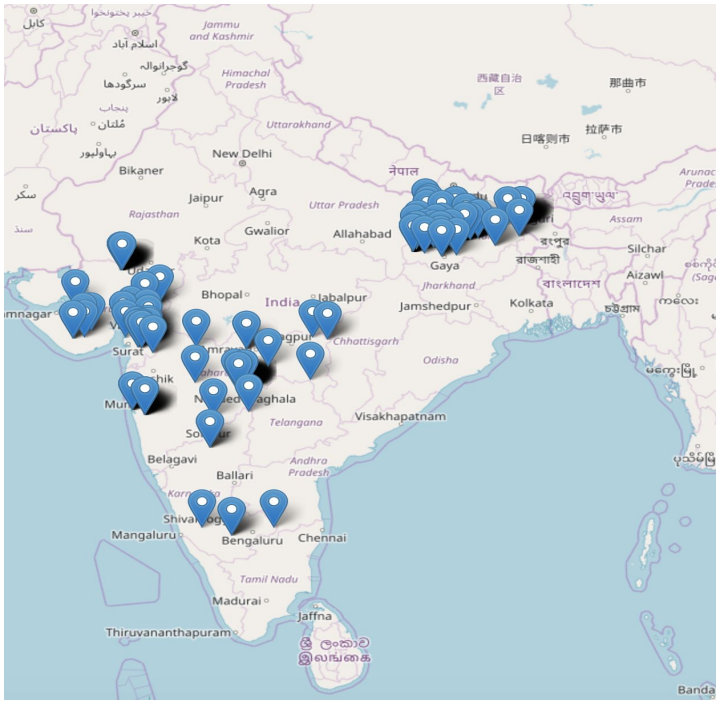



**Figure 06: Instances which support the rule related to event category**

**8.3 Clustering Results.**

Clustering is the process of grouping things which are similar to each other. The similarity measure used here is the distance between two events. Table 04 represents the number of clusters obtained by considering different distances. DBSCAN algorithm is applied on the reduced dataset (10 attributes) with $\varepsilon$ = 0.05 and minimum number of instances in a cluster to 4. Clustering is performed on the dataset corresponding to a single state (Kerala).

**Table 04: Number of clusters and outliers**

| Distance | # Clusters | # Outliers | # Instances in major cluster |
|---|---|---|---|
| Spatial | 24 | 7 | 638 |
| Spatio-Temporal | 96 | 487 | 487 |
| Combined | 96 | 997 | 482 |

The number of clusters formed using just the spatial distance (using the haversine formula) are 24 and the number of outliers are just 7. Figure 07 represents the instances in major cluster formed using spatial distance. We observe that all the events are close in space as we are considering just the spatial distance. The number of clusters formed are also less because most of the events happen in and around cities and since we are considering only the spatial distance most of the clusters will be formed in cities with more number of instances.

The number of clusters formed using both spatial and temporal distances are 96 and the number of outliers are 487. We account the increase in number of clusters to the large time frame we are considering for the purpose of clustering (7 years). Figure 07b represents the instances in major cluster formed using spatio-temporal distance.

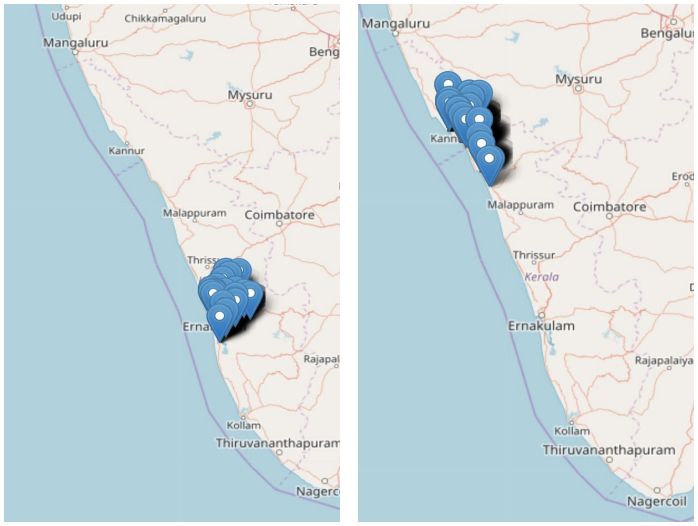
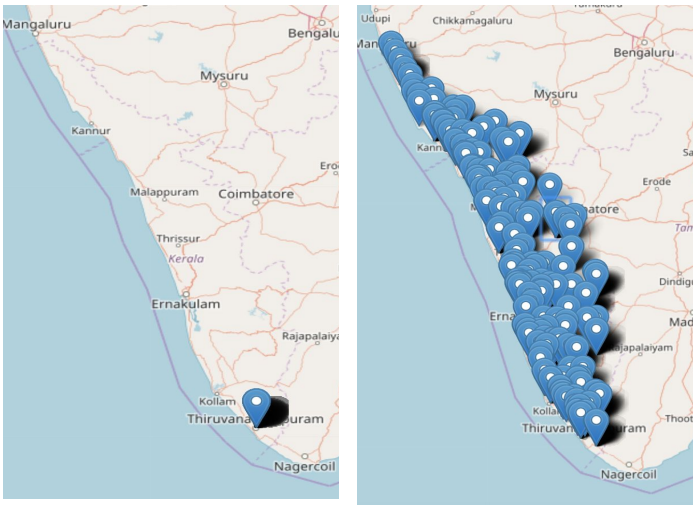



**Figure 07a, 07b: The largest cluster obtained by using spatial distance is shown to the left and the largest cluster obtained using spatio-temporal distance is shown to the right.**

We observe that the events are still clustered together in space as we are giving almost half weight for the spatial distance. Even in this case we observe that most of the events are in and around cities but more spread out compared to the clusters formed using spatial distance.

The number of clusters formed using combined distance are 96 and the number of outliers formed are 997. We observe that there has been an increase in the number of outliers as many events might be occurring and are confined to that specific region. So, those events might be independent in space and time and the socio-economic, infrastructure attributes and hence doesn't belong to any cluster. The major cluster consists of 482 events of which all of them are located at a single location, the capital of the state. As, more number of events occur at the head quarters of a particular state, all of the events might be geocoded at the same location and will have the same socio-economic and infrastructure attributes which accounts for a low distance value and hence clustered together as shown in figures 08a and 08b.




**Figures 08a, 08b: The largest cluster formed by using the combined distance is shown to the left and the outliers obtained in the clustering using combined distance is shown to the right.**

## 9. EVALUATION

The data combined from different sources seems to be accurate as we performed joins of different datasets mainly based on the census id, and if the census id is not available then we combined different datasets using the name of the villages which are first filtered by state and district, sub-district name. The final dataset is then verified for any duplicate rows and are filtered out. We obtained almost 95% accurate dataset containing the correct instances whereas the remaining 5% rows are either duplicate village names in the same sub-district or villages with different names in the shapefile and the census dataset.

For the association analysis, our results are accurate on the reduced dataset and doesn't contain any exceptions where as the rules when interpreted using the original attributes (83) may not be correct in all cases as much of the information is lost due to

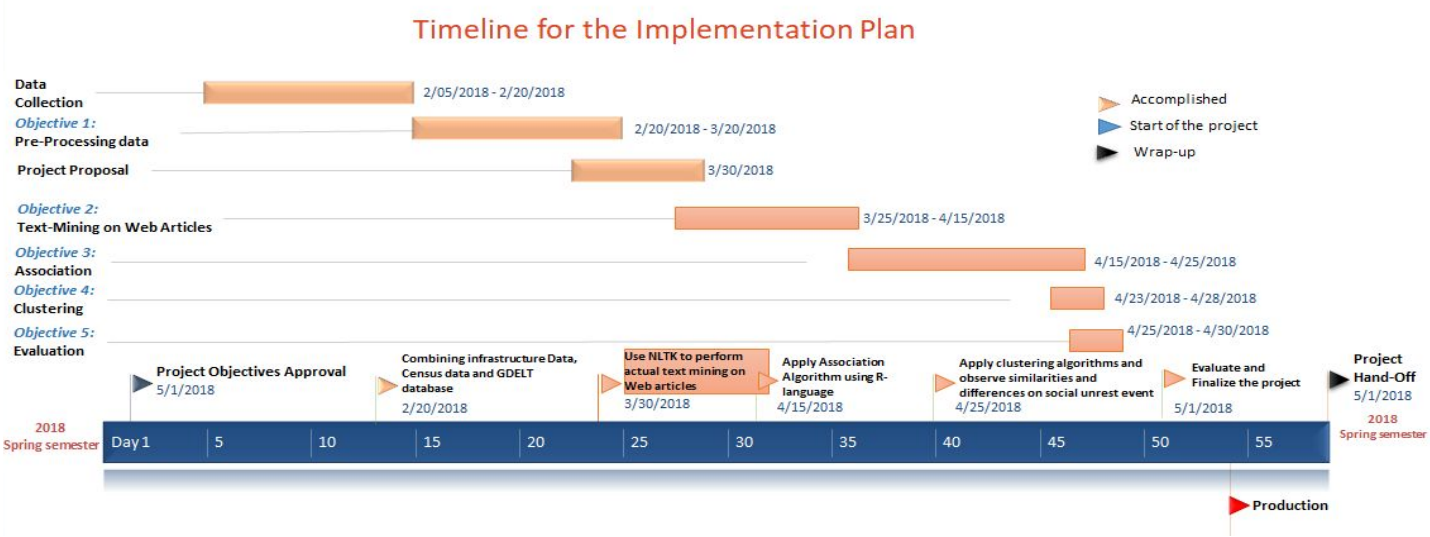



**Figure 9: Implementation Timeline**

dimensionality reduction. However, most of the rules generated are accurate and straightforward to understand.

For clustering, our results are accurate on the reduced dataset and the figures 7, 8 show that the events are also located spatially close to each other if the clustering is done using either spatial or spatio-temporal distances. However, the clustering results may change if the entire dataset (83 attributes) is used.

For the text mining objective, we might have missed some important information while geocoding as most of the place names in India are same as the names of some people. The Geograpy package which is used to obtain the location names from a newspaper article doesn't provide the street names which are usually names of people. To overcome this problem, we used the "other" field reported by the Geograpy package to include the most common names in the location string field. However, we might not geocode the location 100% accurately as the actual event location might vary from the one reported in the news article.

## 10. IMPLEMENTATION PLAN AND TIMELINE

1. The collection and preprocessing of data was completed by February 2018. The cleaning and combining of data was completed by March 2018.
2. A program was written to calculate the distance from a gdelt data point to every infrastructure point. This was completed by March 2018.
3. Web scraping was completed. The NLTK package was used to filter social unrest articles. Then, the Geograpy and Geopy packages were used to obtain the locations for the social unrest events. Google Maps API was used to collect the latitude and longitude points of these events. Finally, the Folium package was used to map and visualize these points. This was finished in April 2018.
4. Association and clustering analysis was performed in the first three weeks of April 2018.
5. All project components were finalized by end of April 2018 and submitted by May 2018.

## 11. CONCLUSION AND FUTURE WORK

In this work, we analyzed the different attributes that contribute to a social unrest event, observed the different events reported in news articles, discovered different clusters of events which are similar to each other either spatially or spatio-temporally, and visualized the different instances supporting an association rule Some future work on this project could include:

- Using a more efficient clustering algorithm to cluster all data
- Completing text mining to obtain the cause of a social unrest event, then clustering events based on the underlying cause
- Using a classifier to classify different protests
- Performing these same analyses to all the states of India or to different countries of interest

## REFERENCES

[1] Aakunuri, Manjula, Dr G. Narasimha, and Sudhakar Katherapaka. "Spatial data mining: a recent survey and new discussions." *International Journal of Computer Science and Information Technologies* 2, no. 4 (2011).

[2] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: linking text sentiment to public opinion time series," in Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM '10), pp. 122–129, Washington, DC, USA, May 2010.

[3] Baeza-Yates, Ricardo. "Searching the future." *SIGIR Workshop MF/IR*. 2005.

[4] Box-Steffensmeier, Janet M., and Bradford S. Jones. *Event history modeling: A guide for social scientists*. Cambridge University Press, 2004.

[5] Braha, Dan. "Global civil unrest: contagion, self-organization, and prediction." *PloS one* 7.10 (2012): e48596.

[6] G. Korkmaz, J. Cadena, C. J. Kuhlman, A. Marathe, A. Vullikanti, and N. Ramakrishnan, “Combining heterogeneous data sources for civil unrest forecasting,” in Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM ’15), pp. 258–265, Paris, France, August 2015.

[7] Hua, Ting, Liang Zhao, Feng Che Chang-Tien Lu, and Naren Ramakrishnan. "How events unfold: spatio-temporal mining in social media." *SIGSPATIAL Special* 7, no. 3 (2016): 19-25

[8] J. Ritterman, M. Osborne, and E. Klein, “Using prediction markets and twitter to predict a swine flu pandemic,” in Proceedings of the 1st International Workshop on Mining Social Media, vol. 9, pp. 9–17, 2009.

[9] K. Leetaru and P. A. Schrodt, “Gdelt: global data on events, location, and tone 1979–2012,” in ISA Annual Convention, vol. 2, Citeseer, 2013.

[10] K. Radinsky and E. Horvitz, “Mining the web to predict future events,” in Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM ’13), pp. 255–264, February 2013.

[11] Kumar, Ch N. Santhosh, V. Sitha Ramulu, K. Sudheer Reddy, Suresh Kotha, and Ch Mohan Kumar. "Spatial data mining using cluster analysis." *International Journal of Computer Science & Information Technology* 4, no. 4 (2012): 71.

[12] M. Arias, A. Arratia, and R. Xuriguera, “Forecasting with twitter data,” ACM Transactions on Intelligent Systems and Technology , vol. 5, no. 1, article 8, 2013.

[13] Qiao, Fengcai, et al. "Predicting social unrest events with hidden Markov models using GDELT." *Discrete Dynamics in Nature and Society* 2017 (2017).

[14] Ramakrishnan, Naren, et al. "'Beating the news' with EMBERS: forecasting civil unrest using open source indicators." *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014.

[15] Renn, Ortwin, Aleksandar Jovanovic, and Regina Schröter. "Social unrest." (2011).

[16] Shin, Dear Sungbok, Minsuk Choi, Jinho Choi, Scott Langevin, Christopher Bethune, Philippe Horne, Nathan Kronenfeld et al. "STExNMF: Spatio-Temporally Exclusive Topic Discovery for Anomalous Event Detection." In *Data Mining (ICDM), 2017 IEEE International Conference on*, pp. 435-444. IEEE, 2017

[17] S. Muthiah, B. Huang, J. Arredondo et al., “Planned protest modeling in news and social media,” in Proceedings of the 29th AAAI Conference on Artificial Intelligence (IAAI '15), pp. 3920–3927, 2015.

[18] S. Muthiah, Forecasting protests by detecting future time mentions in news and social media [M.S. thesis], 2014.

[19] Y. Keneshloo, J. Cadena, G. Korkmaz, and N. Ramakrishnan, “Detecting and forecasting domestic political crises: a graphbased approach,” in Proceedings of the 6th ACM Web Science Conference (WebSci ’14), pp. 192–196, ACM, June 2014.