CSE 635: Web/Text/Social Media Mining

Spring 2019
Instructor: Rohini K. Srihari

Project Description and Requirements

1. Overview

Social media, news, blogs, etc. are rich data sources for predicting and detecting events. Targeted domains could be crime, election, and social unrest. Due to the unstructured language, short-length messages, dynamics, and multi-type of factors that are involved, it is a challenging task to develop a system to forecast or detect relevant events from different data sources using content mining techniques. The goal of this project is to gain hands-on experience in developing practical solutions to societal problems based on web/text/social media mining.

The specific event classes of interest are: global social unrest events, such as demonstrations, marches, and protests. I am especially interested in insights you can provide for how much of this unrest can be attributed to **election violence**. Examples of this type of attribution include (but are not limited to) the visit of a political leader, new policy or scheme announced by a political party, bribing/corruption, dissatisfaction of voting results, intimidation tactics, or accusation of manipulating the election.

This project will satisfy the MS project requirements specified by the CSE department. While the problem definition and evaluation dataset have been fixed, there is ample room for creativity on your part in further enhancement of the solution, and implementation. Be creative, and most importantly pace yourself properly during the semester.

2. Data sources

This project seeks innovative solutions and approaches for discovering and combining multiple data sources. You are encouraged to use all kinds of data sources, historical data, real-time data, etc., in the project as long as you have the rights to use the data. Please cite the corresponding sources in your report properly.

Benchmark dataset

Your system should be evaluated against the benchmark dataset. The Armed Conflict Location & Event Data Project (ACLED) is a disaggregated conflict analysis and crisis mapping project. ACLED is the highest quality, most widely used, real-time data and analysis source on political violence and protest in the developing world. Practitioners, researchers and governments depend on ACLED for the latest reliable information on current conflict and disorder patterns. Refer to https://www.acleddata.com for more information and access to the data.

3. Project tasks: Choose one

Task 1: Predictive analytics

The goal of this task is the automated generation of social unrest related event forecasts. You are encouraged to develop and test innovative forecasting methods that ingest and process publicly available data sources to predict social unrests. By studying and analyzing multiple kinds of data sources, you should also be able to determine the leading indicators/drivers to such events.

The effectiveness of your system should be evaluated in terms of both Lead Time, prediction accuracy, and granularity which may vary depending on country. Lead Time refers to the number of days between the date the forecast was produced and the date the actual event was reported. The Lead Time of your forecasts should be at least **2 days**. You are encouraged to use evaluation metrics that are suitable for this task; multiple evaluation metrics may be used to showcase your results in the best light.

Task 2: Event extraction and summarization

By choosing this task, you are asked to propose and implement a system for extracting and summarizing events in some countries about targeted domains by ingesting multiple input, such as social media and news. Your system should give detailed information about the events so that your users could understand the events by reading the summary. The following information should be provided for each event: event date, location, event type, parties involved, data sources, and a brief description. **Daily based** summarization should be generated on a timely manner, i.e. the delay of your system should be at most **24 hours**.

Requirements (apply to both of the two tasks)

1) Event types are restricted to "**Riots/Protests**" AND "**Violence against civilians**". Detailed definitions can be found in the ACLED Codebook (https://www.acleddata.com/wp-content/uploads/2017/12/ACLED_Codebook_2017FINAL.pdf)
2) Here is a list of counties you should focus on: **India, Indonesia, Thailand.**
3) You need to benchmark your system for at least **1 month**. So, start early!

The following questions should be considered when choosing either of the tasks:

- What is the underlying question that you are trying to answer? Who is your targeted user? What kind of information you need to provide to them?
- How do you intend to answer it? What data sources and features do you think are pertinent?
- How will you model your system and test your hypothesis?
- What should be the evaluation metrics?

4. What to submit

You should plan on preparing for the following:

1) **Project proposal**
   In your proposal, you should define the problem you are trying to solve, your objectives, details on the dataset, how the dataset is processed and adapted by your system, which evaluation metrics are being used, detailed explanation on your proposed system and each component of it, a clear plan of your project – who does what and the targets for each milestone.
2) **Preliminary in-person demo** demonstrating a baseline system implemented
3) **Midterm report** describing initial evaluation and results
4) **Final in-class presentation**
5) **Project report** in conference paper format

5. Grading

- **Milestone 1 (10%):** Project Proposal (week of Feb 25)
  – Present:
  - Project objectives
  - Data set, features to be implemented
  - Evaluation methodology
  - Project plan
- **Milestone 2 (10%):** Baseline results (week of March 25)
  – Show initial demo working (in person/video)
- **Milestone 3 (10%):** Initial Evaluation (week of April 15)
  – Initial evaluation of results (submit report)
- **Milestone 4 (40%):** Final Project Presentation (week of May 6th)
  – In class presentation
  – Project report (KDD paper format) to be submitted
  – All deliverables due by May 2

All project related discussion will be conducted through the piazza site for this course.