Project Deliverable Pt2

Spatiotemporal Analysis and Prediction of Global Social Unrest

GROUP – 20
Anol Kurian Vadakkeparampil 5626-8544
Aadithya Kandeth 6980-2791

Taxonomy (categorization) of Papers:

Category	Definition	Examples
Machine Learning Methods	Approaches that use machine learning algorithms to train predictive models from data.	**[1]** "Predicting Social Unrest Using GDELT" -Divyanshi Galla et al **[2]** "Learning Evolutionary Stages with Hidden Semi-Markov Model for Predicting Social Unrest Events" - Fengcai Qiao et al **[3]** "Predicting Social Unrest using Sentiment Analysis" - Ufuk Ozdemir **[4]** "'Beating the News' with EMBERS: Forecasting Civil Unrest using Open Source Indicators" - Naren Ramakrishnan et al **[5]** "Forecasting Civil Unrest Using Social Media and Protest Participation Theory" - Congyu Wu et al, **[6]** "Civil Unrest Prediction: A Tumblr-Based Exploration" - Jiejun Xu et al
Agent-Based Models	Approaches that use agent-based modeling to simulate the behavior of complex systems, such as social systems.	**[7]** "Entity-Based Integration Framework on Social Unrest Event Detection in Social Media" - Ao Shen et al
Statistical Methods	Approaches that use statistical methods to identify patterns and correlations in data to develop predictive models.	**[8]** "Global Civil Unrest: Contagion, Self-Organization, and Prediction" - Dan Braha
Multi-Source Models	Approaches that combine data from multiple sources to improve the accuracy of predictive models.	**[9]** "Multi-source models for civil unrest forecasting" - Gizem Korkmaz et al ****[4]** "'Beating the News' with EMBERS: Forecasting Civil Unrest using Open Source Indicators" - Naren Ramakrishnan et al

Categorizations of Papers are not mutually exclusive, they can belong to multiple categories;* ***[4] *is an example; Additional Categorization can also be based on data source; Due to space constraints they are omitted.*

Breakdown for each method in category:

Paper	Major Steps	Advantages	Limitations
[1]	▪ Extract economic, political, and social features from GDELT event database. ▪ Train regression models to correlate features with historical instances of unrest. ▪ Apply models to forecast unrest likelihood by country.	▪ Quantifies public opinion/mood. ▪ Model predicted >80% of unrest incidents. ▪ Real-time sentiment could help authorities prepare.	▪ Focuses only on the US and its counties (not global). ▪ Validation performed manually against news articles could introduce media biases and miss less publicized events over reliance on media.
[2]	▪ Identify and label unrest events from GDELT using keyword filtering ▪ Construct sequential unrest event graph. ▪ Train hidden semi-Markov model to learn typical event sequencing patterns. ▪ Apply model to predict most likely future event at country level.	▪ Models evolving stages & topics of unrest. ▪ Integrates news & protest data. ▪ Performed better than baselines. ▪ Provides insights into typical progressions.	▪ Focuses only 5 South Asian countries so may not generalize well to other countries without retraining. ▪ External factors like police actions and government responses are not captured.
[3]	▪ Collect labeled corpus of social media posts related to historical unrest events. ▪ Extract text sentiment features ▪ Train classifier models to categorize unrest-related posts. ▪ Apply classifier to new posts to estimate level of emerging unrest risk.	▪ Can capture the sentiment of social unrest events, which is an important indicator of the potential for social unrest. Can be used to make predictions without the assumption that past patterns will continue into the future.	▪ Sentiment analysis has challenges like sarcasm, ambiguous language so advanced NLP is needed. ▪ Not everyone uses social media, so sample population is limited. ▪ External factors like police action also drive unrest, not just public sentiment.
[4]	▪ Compile diverse dataset of news, social media posts, economic indicators. ▪ Extract features including sentiment, socioeconomic metrics, political events. ▪ Train ensemble model to predict civil unrest. ▪ Apply model to forecast unrest likelihood for different countries.	▪ Uses diverse online indicators. ▪ Data-driven without hand-coded rules. ▪ Forecasts events 7 days in advance decently. ▪ Provides insights into important precursors.	▪ Accuracy shows a significant decline for predicting events beyond 7 days. ▪ Requires significant computational resources for processing massive data due to the utilization of over 5 models to calculate cumulative event probabilities. ▪ Considers only text sources so other events could be missed.
[5]	▪ Gather geo-located tweets and identify protest-related posts.	▪ Incorporates social science theory. ▪ Provides interpretable inputs based on participation factors.	▪ Data collection was focused on bigger protests covered by media.

	▪ Extract user profile features like demographics and network connections. ▪ Build regression model to explain user engagement. ▪ Apply model to estimate size and locations of emerging protests.	▪ Forecast protest size a month in advance fairly accurately. ▪ Better performance than baselines.	▪ Assumes participation theory fully explains protest mobilization. ▪ External influences on protests not captured in indicators.
[6]	▪ Collect dataset of Tumblr posts related to civil unrest. ▪ Extract post features including text, reblogs, user demographics, and metadata. ▪ Train classifier models to categorize unrest-related posts. ▪ Apply classifier to new posts to estimate level of emerging unrest risk.	▪ Novel Tumblr data source. ▪ Combined text, image, network indicators. ▪ Provided insights into predictive indicators. ▪ Decent baseline prediction performance.	▪ Single case study focused only on Ferguson unrest. ▪ Tumblr data likely suffers from sample bias compared to full population
[7]	▪ Identify civil unrest-related tweets using keyword filtering and hand-labeled data. ▪ Extract entities like people, locations ▪ Construct graph linking co-occurring entities. ▪ Apply graph neural network model to detect unrest-related communities.	▪ Entity information provides useful contextual signals. ▪ Entity clustering allows grouping related terms. ▪ Better precision & recall than baselines. ▪ Can detect localized events.	▪ Assumes social unrest is limited to protests, while it encompasses various forms such as migration and inflation. ▪ Only looks at entities in tweets. Does not use other information like emotions, images, connections between users. ▪ Works better for detecting ongoing protests and not predicting future protests.
[8]	▪ Develop agent-based model representing individuals, governments, and other entities. ▪ Set rules for agent actions like protesting. ▪ Run simulations to study unrest emergence dynamics. ▪ Analyze results to determine factors that make large-scale unrest more likely.	▪ Global view of interconnections & cascading effects. ▪ Agent-based modeling of unrest spread. ▪ Understanding of local to global propagation. ▪ Elucidates role of factors like information flows.	▪ Looks only at the geographical network and telecommunication connections between countries to analyze social unrest. ▪ Addresses high-level prediction but lacks operational contagion model and protest data forecasting specifics.
[9]	▪ Gather diverse data streams on past unrest. ▪ Process data into common format and fuse into consolidated dataset ▪ Extract relevant features. ▪ Build ensemble machine learning model to forecast unrest incidents.	▪ Combines complementary data sources. ▪ Allows varied indicators. ▪ Ensemble model outperformed individual models. ▪ Provides framework to add more sources.	▪ Focused on predicting general likelihood, not specific event details. ▪ Only tested in Venezuela so limited geographical focus. ▪ Requires access to extensive high quality data sources

---

Remaining challenges of the problem:

1. Generalization Across Regions and Cultures:
    a. Some papers do not explicitly account for the variations in cultural, political, and social contexts when making predictions. Addressing generalization across diverse regions and cultures remains a challenge.
    b. Correlation-Causation: While studies in the field can provide a probability of an event occurring, they are unable to clearly explain why this happens and the underlying reasoning and causation of global scenarios. Current methods simply correlate unrest with social factors while potential solutions could employ causal inference to identify factors that actually cause social unrest.
2. Geographical Diversity:
   Most of the papers focused on a singular geographical location (USA/ European Union) and hence failed to make use of complex inter-country geopolitical relationships. This is a major gap in predicting social unrest globally.
3. Multi-source data fusion:
   Combining data from multiple sources can improve the accuracy of predictive models. However, it can be computationally expensive to fuse and analyze data from multiple sources. Finding the right line between the computation power/ prediction accuracy trade-off is a challenge.
4. Assumptions about data:
    a. Almost all approaches assume that future events would follow historical patterns, but that is not often the case.
    b. Most data sources are assumed to be ethically clean and non-biased. Social media and news sources can be a subject to political or financial influences. It is also difficult to predict the impact of different interventions on social unrest.
5. Real-time Prediction:
   The ability to make real-time predictions is not fully addressed in a majority of the literature. Developing real-time prediction models is a notable challenge for applications requiring timely responses.