Text Analysis of News Articles (Building a Protest Dataset through Machine Learning)

Howard Liu (hao.liu@duke.edu) Sophie Lee (sophie.lee@duke.edu), & Haohan Chen (haohan.chen@duke.edu)

Department of Political Science, Duke University

Objectives

In this project, we attempt to build a dataset of political conflict. We aim to extract important information with machine learning algorithm from news reports of political conflicts. Specifically, our methods automatically code entries of news on: *where* the conflicts happen; *who* are involved, and; *what* type of issues are at stake.

We attempt to improve on the *state-of-art* of the political science scholarship by tackling the following challenges:

- Separating multiple events in one news report
- Machine-coding actors involved in conflicts
- Machine-coding types of issues at stake

Introduction

Building a reliable data is crucial in political science as political scientists study phenomena involving human interactions that are not always clean-cut. Although domestic protest is one of the central interest of peace and security research, building a comprehensive data that encompass all countries is extremely difficult due to challenges non-democratic regimes pose on researchers collecting data. Hence, analyzing news articles has become a new way of studying many aspects in such societies, yet handling massive texts and producing a dataset has still room for improvement. One existing data, that is widely used in the field is World-Wide Integrated Crisis Early Warning System (W-ICEWS)[1], yet the data suffer from the following three problems.

First, ICEWS does not account for news articles that contain more than one events.

Second, ICEWS does not differentiate location identifiers that refer to the location of news agencies and the location of the events.

Third, ICEWS does not code the event type.

Hence, we aim to resolve these issues and improve the accuracy of the current ICEWS data.

Data

We train and test our model with a subset of ICEWS data containing 642 English news reports from 2001 to 2014 on protests in China. Variables include:

- Original text of the report
- Human-coded location
- Human-coded issue types

For the the coding of location and actors, we fitted the model with the full dataset. For issue types, 80% of the cases were used as the training set and the rest as the test set.

Frequency of Protests in Each Chinese Province (2001-2014)

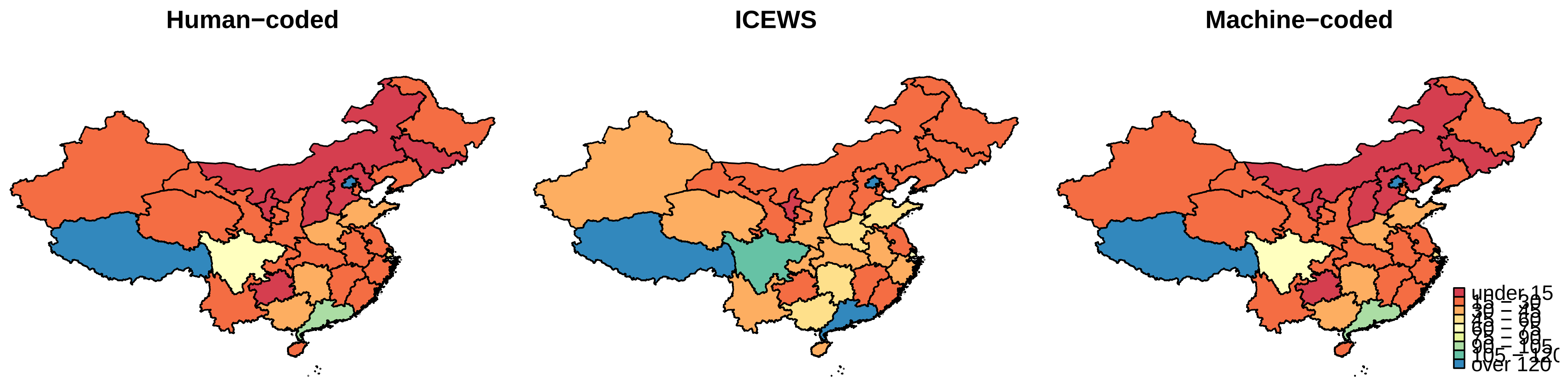

Figure 1: Figure caption

Important Results

- Locations are coded correctly for 66% of the cases, which is substantially higher than the current ICEWS data, of which accuracy rate is 47%.
- We were able to code actors and issue types as well, yet (1) we do not have a manually coded variable for the former and hence unable to compute the accuracy rate, and (2) the performance of the latter still needs to be improved greatly.

Methods

We first identified the news sources that contaminate the data and removed lines that are not part of the news articles. We then used a dictionary of provinces and search for province names. Based on the number of provinces in one article, we analyzed the text using the Latent Dirichlet Allocation (LDA) method, setting the number of topics equal to the number of provinces produced. Then, we examined the most frequent words of each topic and searched for actor types.

To code issue types, we first created the document term matrix. We then used the human-coded issue type as the outcome and fit the dtm into SVM, Bagging, Boosting, Random Forest, and Trees. We compared the performance and examined consensus rates of each model. [2].

Performance

The performance substantially improve upon existing methods in geo-coding conflict events. Nevertheless, our attempt on actor and event type coding, albeit its originality, has less-than-ideal performance. We report the accuracy rate of matching coding, using human coding and ICEWS data as benchmarks.

Data	Manual	ICEWS	Ours
Location	100%	47%	66%
Actors	missing	100%	3%

Table 1: Accuracy rates: Location and Actors

Model	Precision	Recall	F-Score
SVM	.0267	.111	.043
Boosting	.067	.072	.064
Bagging	.129	.121	.104
Random Forest	.043	.068	.044
Tree	.104	.088	.081

Table 2: Accuracy rates: Issue Type

Concluding Remarks

- The machine-coding methods we develop outperform the *state-of-art* in geo-coding.
- Our methods have potential to code and classify events on more characters that previously rely on human coding.

Limitation and Future Works

- The training set is small, thus plan to expand it.
- The performance of actors and issue types need much improvement. We plan to further examine the sparsity of DTM, classification model, and pre-processing with post-tagging.

References

[1] Martin, Lockheed. 2014. Integrated Crisis Early Warning System (ICEWS).

[2] Jurka, Timothy P and Collingwood, Loren and Boydstun, Amber E and Grossman, Emiliano and Others. 2011. Rtexttools: A supervised learning package for text classification