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

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

Department of Political Science, Duke University

December 8, 2015

1 Objective

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

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

- Separating multiple events in one news report
- Improving the accuracy of locations in the data.
- Machine-coding actors involved in conflicts
- Machine-coding types of issues at stake

2 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 political protests is one of the central interests in 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]. Despite its wide applications, the data suffer from the following three problems that we aim to fix.
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.

3 Motivation

Importance of the news sources and protest data.

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 many non-democratic regimes pose on researchers collecting data. Hence, analyzing news articles in the world has been a new way of analyzing many aspects in such societies, yet handling a massive texts and producing a dataset has room for improvement. One existing data, that is widely used in the field is World-Wide Integrated Crisis Early Warning System (W-ICEWS), yet the data suffer from the following two 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 type of event.

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

4 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.

5 Method

We first identified the news sources that contaminated the data and removed lines that are not part of the news articles. For instance, many raw data, the original news articles, contain the information of the news agencies at the beginning or at the end in the format of "location of the news agency, name of the news agency, and the date or publication." We created a character vector containing news agency names and searched for phrases that 1) contain those news agency names and 2) are located either at the beginning or end or each article. Once identified as source indicators, they were removed accordingly.

Next, we created a province names and searched for those province names. In this case, we used 34 Chinese province names as the dictionary. As the final step of cleaning the text, we extracted stem words and lowered the cases. Based on the number of provinces in one article, we analyzed the texts using the Latent Dirichlet allocation (LDA) method, setting the number of topics equal to the number of provinces produced. This method was chosen because many news articles report more than one event. For instance, one of the articles report a protest case in Guizhou province, another one in Henan. Yet, the ICEWS dataset codes this article as one event that occured in Beijing, the location of the news agency. Hence, after cutting out the irrelevant information, our algorithm split the events based on the number of locations.

We then examined 20 most frequent words of each topic and searched for actor types in the same way that we searched for the province names. Finally, the locations and the actor types extracted from our algorithm are compared to the manually coded data.

6 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.

7 Performance

Assuming the manually coded data are 100% correct, we computed the accuracy rates of the current ICEWS data as well as the data generated by our algorithm. (Caveat: We weren't able to code the actors variable in time for this project, but we intend to code the actors manually and compare ICEWS and our data in the future.)

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

Table 1: Accuracy rates

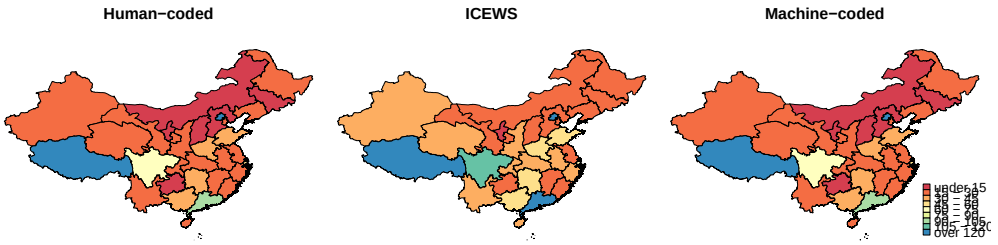

Figure 1: Frequency of Protests in Each Chinese Province (2001-2014)

For the location data, our algorithm performs about 20% better than the current ICEWS data.

8 Limitations and Future Works

- The training set is small, thus we 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.

9 Concluding Remarks

- The machine-coding methods we developed outperform the current ICEWS data in geo-coding.

- By analyzing news articles in the world, our methods have potential to code and classify events that previously relied on human coding.

10 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

[3] R package:`topicmodels` by Grun, Bettina and Hornik, Kurt. Jul. 2015.