CMU Civil Unrest Prediction

David Bamman

The prediction model described here is composed of four binary logistic regression classifers – one for each combination of $\{es, pt\}$ and $\{city, country\}$ – each with a sparse ℓ_1 regularizer, trained to minimize the empirical risk of data from May 1 – September 30, 2012. Each model is trained on all locations within that 4-way distinction, with fixed effects for the identity of each location, and uses data from day t to predict whether an event will take place for each of those locations at time $t + 1$.

The target locations include all 21 countries in Latin American + 54 cities (top 50 cities by population + tegucigalpa, la paz, panama city, san jose [which had more than 10 events in 2012]).

0.1 es_country

argentina
belize
bolivia
chile
colombia
costa_rica
ecuador
el_salvador
french_guyane
guatemala
guyana
honduras
mexico
nicaragua
panama
paraguay
peru
suriname
uruguay
venezuela

0.2 pt_country

brasil

0.3 es_city

asuncion paraguay
barranquilla colombia
bogota colombia
buenos_aires argentina
cali colombia
caracas venezuela
cartagena colombia
cordoba argentina
ecatepec mexico
guadalajara mexico
guatemala_city guatemala
guayaquil ecuador
iztapalapa mexico
juarez mexico
la_paz bolivia
leon mexico
lima peru
madero mexico
managua nicaragua
maracaibo venezuela
maracay venezuela
medellin colombia
mexico_city mexico
monterrey mexico
montevideo uruguay
nezahualcoyotl mexico
panama_city panama
quito ecuador
rosario argentina
san_jose costa_rica
santa_cruz_de_la_sierra
santiago chile
tegucigalpa honduras
tijuana mexico
valencia venezuela
zapopan mexico
zaragoza mexico

0.4 pt_city

belem brasil
belo_horizonte brasil
brasilia brasil
campinas brasil
curitiba brasil
fortaleza brasil
goiania brasil
guarulhos brasil
iguacu brasil
maceio brasil
manaus brasil
porto_alegre brasil
recife brasil
rio brasil
salvador brasil
sao_luis brasil
sao_paolo brasil

1 Features

1.1 Fixed effects

All fixed effects take a dummy value of 1.

Location x Month The identity of the location crossed with the identity of the month (e.g., *argentina_7*). Separating out the location-specific bias terms from the month in which we observe data is crucial to avoid letting early months (in which there is little reliable training data) exert too much influence on the learned biases. In the current model, biases are learned for each location x country, and at prediction time, the most recent month is used for the location effect.

Day of Week The day of the week (1-7). Including this feature originated out of inspecting the learned weights from early versions of the unigram model (see random effects below), where words like *miércoles* (Wednesday) were found to be good predictors of events. Formalizing this as a fixed effect leads to a substantial improvement over the baseline – events very rarely occur on the weekend.

1.2 Random effects

All random effects take three values: 0 if unseen in all tweets on day d in location p; 1 if present but occuring less frequently than 2 standard deviations above the mean frequency for that feature across all locations and days; 2 if occuring more frequently than 2 s.d. Note: I experimented with binary values (1 if present, 0 if not), normalized frequencies, raw counts, log counts, log frequencies, and frequencies normalized to z-score (subtract mean, divide by s.d.); this setting performed best on development data.

Hand. 168 words (*ocupa*) and phrases (*va a marchar*) drawn from a manually curated list (`Dropbox/OSI-Shared/BBN Data/allqueryterms.txt`).

HandTom. All terms in HAND conjoined with manana/mañana (es.) or amanha/amanhã (pt.)

Unigram. The most frequent 25,000 unigrams in the complete dataset.

UnigramTom. The most frequent 25,000 unigrams conjoined with manana/mañana (es.) or amanha/amanhã (pt.)

2 Evaluation

To evaluate our models, we train on data from May 1 – August 31, 2012 and test on data from September 1 – September 30. Over the test time period, there are 159 events in Spanish-speaking countries (es_country) and 88 events in Spanish-speaking cities (es_city). Brasil (with 5 events) and Brazilian cities (with 4 events) are excluded from evaluation due to the small sample size.

In the es_country evaluation, including day-of-week fixed effects has the largest improvement on accuracy; lexical features (including the hand-curated features and those drawn from the top 25,000 unigrams overall), have a slight but not statistically signficant effect. For es_city, the best strategy is to always choose no event.

Feature	MSE	0-1 Accuracy	# Correct
Location Baseline	0.179	0.735	441
Location x Month	0.181	0.752	451
+ Day of Week	0.171	0.765	459
++ HandTom/Hand	0.171	0.765	459
+++ UnigramTom	0.171	0.770	462
++++Unigram	0.171	0.767	461

Table 1: es_country accuracy.

Feature	MSE	0-1 Accuracy	# Correct
Location Baseline	0.061	0.926	1028
Location x Month	0.062	0.917	1018
+ Day of Week	0.060	0.920	1021
++ HandTom/Hand	0.059	0.922	1023
+++ UnigramTom	0.061	0.920	1021
++++Unigram	0.063	0.918	1019

Table 2: es_city accuracy. Always guess no event performs best.

These numbers are slightly misleading, however, due to the overwhelming dominance of no-event days (both at the country and especially at the city level). If our task is to predict as many true events as possible, at varying degrees of acceptable false positive rates, a better evaluation is the ROC curve. Figure 1 plots this curve for the 20 Spanish-speaking countries, illustrating the effect of varying the acceptance threshold in the binary classification. Each of the 100 data points represents a choice of threshold in the range [0,100]; if the confidence of the classifier is above this threshold for some particular input, that instance is labeled as an event (i.e., "positive"). The red point at (.33, .68), for example, corresponds to a threshold of .16; while this threshold has a false positive rate of 33%, it captures 68% of the true events. A more stringent threshold of .70 has false positive rate of 1.5% while capturing 11.3% of true events.

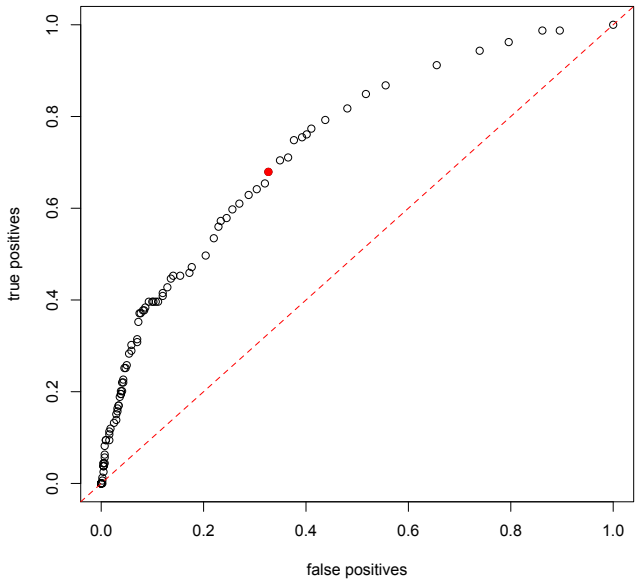

Figure 1: ROC curve for es_country under the +++UnigramTom model. The red point corresponds to a false positive rate of 33% and true positive rate of 68%; the dashed red line indicates chance performance, with all points to the left of the line performing better than chance.

3 Dominant Features

0.31048799221325135	movilización	unigramTom
0.2711484041882385	escuela	unigramTom
0.2617738890295068	guardare	unigramTom
0.2528731154817262	damos	unigramTom
0.24992737386636296	muchísima	unigramTom
0.24105855952885977	tuit	unigramTom
0.2311606432399665	colapso	unigramTom
0.2190068095697447	convocatoria	unigramTom
0.21578936626852896	@udechile	unigramTom
0.21452788600616238	importarle	unigramTom
0.21426082688193818	besan	unigramTom
0.20025101245430552	volver	unigramTom
0.18799244924144623	:-)	unigramTom
0.1862571733393598	agradecemos	unigramTom
0.18286091707089333	segundo	unigramTom
0.1785254079491763	temas	unigramTom
0.17152942444984604	despido	unigramTom
0.1560867585245553	fuego	unigramTom
0.15248799163492854	activo	unigramTom
0.1518391918565881	anuncia	unigramTom
0.151660724133199	kon	unigramTom
0.15056405089207153	va_a_ocupar	hand
0.14526469130458705	glee	unigramTom
0.14281034271787407	màs	unigramTom
0.13277426588827834	copio	unigramTom
0.1301364770800818	dice	unigramTom
0.12799427305443842	huesos	unigramTom
0.12721168440498815	cinco	unigramTom
0.1251903352247223	protestas	unigramTom

Table 3: Highest weights among random effects in es_country model, trained on all available data from May 1, 2012 – September 30, 2012. Amid noise, *movilización* (mobilization), *colapso* (stoppage), *convocatoria* (announcement), *anuncia* (announce), *va a ocupar* (go occupy) and *protestas* (protest) all have face validity.

4 Pre-fab models

From its performance on dev data and the face validity of at least some of the learned feature weights, I've included pre-trained models for the +++ UnigramTom feature set, which includes fixed effects for location x month and day of week and random effects for hand-curated terms and the unigrams conjoined with translations of *tomorrow*). While the results shown above are trained on data from May 1, 2012 – August 31, 2012 and tested on data from Sept. 1, 2012 – September 30, 2012, these models are trained on all available data, from May 1– September 30.

5 Observations

5.1 Nature of the problem

Social media data should be able to predict strikes, occupations and other such events since they are often endogenous to the medium – originating within and disseminating through the social network structure that we can observe. First, this assumption is valid for events involving individuals actively engaged in social media (e.g., "Chilean university students protest for higher budgets," Valparaiso, Chile, 5/23/12), but naturally breaks down completely for individuals who are not (e.g., "Argentina farmers in grain export quota strike," Mexico 5/13/12). Second, efforts like the present study that attempt to exploit the large-scale language model associated with a particular region may be using too coarse of an instrument: it may not be

reasonable to assume that the textual effects will be noticable unless describing an event that is reasonably large in scope (such as a Tahrir square-level event). A more fruitful line of pursuit may be to focus on finding the needles in the haystack – the sources of information whose tweets are reliable indicators of future events.

5.2 Weather fixed effects

Sunny also emerged as a high-ranking feature for several of the development models, suggesting that events may be correlated with the weather, with more events occuring on sunny days than on days with rain. Note that in making a prediction with information at time t for time $t + 1$, weather information at time t may still be relevant, since it can effect the planning for events the next day.

5.3 Markovian temporal effects

One modeling possiblity that we considered but did not pursue was adding parameters to reflect the latent event state for a given location, to help model the probablity of transitioning from a non-event state to an event one, or (more likely) to make an event at time $t + 1$ more likely if there were an event at time t (modeling event momentum). We discarded this idea after noting that the probabilty of an event in a given location at time $t + 1$ in the training data was roughly independent of an event there at time t.

6 Geography

Tables 4 and 5 list the median number of tweets associated with each target location for each day between May 1, 2012 – September 30, 2012 from the Twitter Gardenhose, using the geolocation method described below (see `files/locationTweetCounts` for daily figures).

brasil	400119
mexico	222484
argentina	186429
venezuela	175739
colombia	132133
chile	75299
ecuador	44378
guatemala	42569
paraguay	27652
peru	25005
el_salvador	20149
honduras	10802
nicaragua	9439
panama	9090
uruguay	8796
costa_rica	5322
bolivia	1469
belize	28
french_guyane	3
guyana	2
suriname	2

Table 4: Median number of daily tweets by country over the period May 1, 2012 – September 30, 2012.

sao_paolo	69040
buenos_aires	47659
mexico_city	43232
rio	41960
bogota	33024
caracas	28814
santiago	25868
barranquilla	19023
porto_alegre	18679
guayaquil	16447
curitiba	15124
lima	14227
valencia	13762
monterrey	13368
maracaibo	11556
belo_horizonte	9670
medellin	9570
asuncion	8943
guadalajara	8786
manaus	8596
fortaleza	8459
recife	8422
cordoba	7163
rosario	7104
salvador	6738
cali	6427
maracay	5871
cartagena	5150
brasilia	3972
guatemala_city	3712
quito	3590
montevideo	3575
campinas	3573
guarulhos	2897
tegucigalpa	2773
leon	2185
maceio	2139
tijuana	1507
managua	1249
goiania	1193
juarez	1038
panama_city	1009
zapopan	730
belem	606
san_jose	431
la_paz	392
ecatepec	336
madero	320
zaragoza	165
iguacu	143
sao_luis	109
nezahualcoyotl	97
iztapalapa	84
santa_cruz_de_la_sierra	82

Table 5: Median number of daily tweets by city over the period May 1, 2012 – September 30, 2012.

6.1 Geolocation

Rather than relying on sporadic geo-location info embedded in tweets, we associate tweets with particular regions by matching the user-declared location field to high-precision gazetter drawn from a list of user-specified locations, ranked by the number of unique users with that location string in all tweets from September 2011 through August 2012, discarding ambiguous locations (such as *san jose* or *santigo buenos aires*). The following lists a sample of most frequent user location strings associated with several cities (full list can be found in `files/cities.txt`).

```
66374 santiago chile
42956 santiago
30833 santiago de chile
4420 chile santiago
346 santiago centro

72386 caracas
69371 caracas venezuela
4558 venezuela caracas
1190 caracas vzla
401 caracas ve
321 caracas_venezuela
307 caracas city
205 caracas valencia esp
155 gran caracas
149 en caracas
144 caracas distrito capital
134 caracas dtto capital
113 caracas vnzla

82369 buenos aires
56835 buenos aires argentina
9218 argentina buenos aires
1505 ciudad de buenos aires
770 buenos aires capital federal
613 buenos aires arg
586 general rodríguez buenos aires
509 capital federal buenos aires
291 buenos aires capital
5151 bs as argentina
2411 argentina bs as
1658 capital federal argentina
6615 bs as
5170 capital federal
770 buenos aires capital federal
736 argentina capital federal



42877 mexico city
39506 mexico df
37849 mexico d f
35197 méxico d f
15681 méxico df
10288 ciudad de méxico
5664 méxico city
4664 ciudad de mexico
3780 méxico distrito federal
3325 mexico distrito federal
2993 df mexico
2222 ciudad de mexico mexico
1820 distrito federal méxico
1664 distrito federal mexico
1621 d f mexico
1474 df méxico

10748 la paz
3422 la paz bolivia

10409 managua
3504 managua nicaragua
125 nicaragua managua
46 managua nic
14 managua nicargua

11965 tegucigalpa
6944 tegucigalpa honduras
753 honduras tegucigalpa
83 tegucigalpa hn
74 tegucigalpa honduras c a
34 tegucigalpa m d c
24 tegucigalpa m d c honduras
21 tegucigalpa hnd
20 tegucigalpa honduras ca
18 tegucigalpa hoduras
15 tegucigalpa mdc
14 tegucigalpa_honduras
13 tegucigalpa francisco morazan
12 tegucigalpa | honduras

14707 asuncion
8114 asuncion paraguay
7568 asunción paraguay
2783 asunción
597 paraguay asuncion
450 asuncion py
259 paraguay asunción
210 asunción py
111 asuncion_paraguay
74 asunciòn paraguay
49 asuncion del paraguay
45 asunciòn
42 asunción del paraguay
38 la asuncion
38 asunción_paraguay
27 asuncio paraguay

2786 san jose costa rica
115 costa rica san jose

25793 guadalajara
```

10701 guadalajara jalisco
6467 guadalajara mexico
4628 guadalajara jal
2759 guadalajara jalisco méxico
2563 guadalajara méxico
2438 guadalajara jalisco mexico
1085 guadalajara mx
665 jalisco guadalajara
555 guadalajara jal mexico
463 guadalajara mex
437 guadalajara jal méxico
358 guadalajara jal mex
225 mexico guadalajara
160 guadalajara jalisco mex
132 guadalajara jalisco mx
102 guadalajara jal mx
60 mexico guadalajara jalisco

1769 panama city panama
725 panamá panamá
620 panamá city
575 ciudad de panamá
210 ciudad de panamá panamá
142 panamá city panamá

33728 monterrey
11554 monterrey n l
8306 monterrey mexico
5814 monterrey nuevo leon
4674 monterrey nl
4018 monterrey méxico
3255 monterrey nuevo león
1917 monterrey n l mexico
1874 monterrey nuevo leon mexico
1592 monterrey mx
1490 monterrey nuevo león méxico
1291 monterrey n l méxico
1009 monterrey nl mexico
414 monterrey mex
381 monterrey nl méxico
327 mexico monterrey
287 monterrey nl mex
216 monterrey n l mex
171 monterrey nuevo leon méxico
161 monterrey nuevo león mexico
156 en monterrey
151 monterrey nuevoleon
133 monterrey nl mx
124 monterrey city
117 monterrey n l mx
92 mexico monterrey n l
86 monterrey nuevo leòn
74 monterrey mèxico
68 mexico monterrey nuevo leon
61 méxico monterrey

16484 montevideo
12821 montevideo uruguay
1023 uruguay montevideo
100 montevideo uy

18537 maracaibo
16934 maracaibo venezuela
3133 venezuela maracaibo
2633 maracaibo edo zulia
1746 maracaibo zulia
1067 maracaibo estado zulia
1047 maracaibo zulia venezuela
628 zulia maracaibo
436 maracaibo vzla
430 venezuela zulia maracaibo
299 maracaibo edo zulia venezuela
173 venezuela maracaibo edo zulia
158 maracaibo city
114 venezuela maracaibo zulia
103 maracaibo_venezuela

37945 lima
32734 lima peru
30839 lima perú
3347 peru lima
1490 perú lima
828 lima perù

We construct a country-level gazetter analogously. All location names occurring over 10 times in 2011-2012 data containing the name of the country + the top 20 unique locations from this list with the country name stripped (e.g., "buenos aires"), manually filtered for cleanness (e.g., "new mexico," "san jose" removed for ambiguity with CA city). The following shows a sample (full data in `files/countries.txt`).

```
==> argentina.txt <==
220608 argentina
56835 buenos aires argentina
9218 argentina buenos aires
8264 cordoba argentina
8199 córdoba argentina
6649 mendoza argentina
5586 rosario argentina
5151 bs as argentina
3452 salta argentina
3140 santa fe argentina

==> bolivia.txt <==
7721 bolivia
3422 la paz bolivia
2918 santa cruz bolivia
1080 cochabamba bolivia
614 montero bolivia
590 camiri bolivia
309 tarija bolivia
289 sucre bolivia
246 bolivia santa cruz
121 oruro bolivia

==> chile.txt <==
146030 chile
66374 santiago chile
30833 santiago de chile
5446 concepción chile
5218 viña del mar chile
4537 temuco chile
4458 antofagasta chile
4420 chile santiago
3160 concepcion chile
3053 iquique chile

==> cr.txt <==
24681 costa rica
2786 san jose costa rica
2447 san josé costa rica
959 heredia costa rica
656 cartago costa rica
556 alajuela costa rica
130 puntarenas costa rica
125 guanacaste costa rica
115 costa rica san jose
96 costa rica ms

==> mex.txt <==
268558 mexico
112646 méxico
42877 mexico city
39506 mexico df
37849 mexico d f
35197 méxico d f
15681 méxico df
10288 ciudad de méxico
8306 monterrey mexico
7582 estado de méxico

==> nic.txt <==
9122 nicaragua
3504 managua nicaragua
249 leon nicaragua
206 granada nicaragua
158 esteli nicaragua
141 león nicaragua
140 chinandega nicaragua
131 masaya nicaragua
125 nicaragua managua
116 matagalpa nicaragua

==> panama.txt <==
38878 panama
3326 panama city
1769 panama city panama
1175 panama panama
412 ciudad de panama
342 panama colon
338 colon panama
335 panama city beach
230 chiriqui panama
212 panama chiriqui

==> paraguay.txt <==
39353 paraguay
8114 asuncion paraguay
7568 asunción paraguay
1073 ciudad del este paraguay
1051 san lorenzo paraguay
959 luque paraguay
918 fernando de la mora paraguay
706 concepción paraguay
625 itauguá paraguay
607 caapucú paraguay

==> peru.txt <==
37538 peru
32734 lima peru
30839 lima perú
21629 perú
3347 peru lima
2007 trujillo perú
1714 trujillo peru
1490 perú lima
1405 arequipa peru
1089 arequipa perú
```

```
==> uruguay.txt <==
24410 uruguay
12821 montevideo uruguay
1023 uruguay montevideo
621 salto uruguay
608 maldonado uruguay
472 punta del este uruguay
364 canelones uruguay
272 rivera uruguay
252 florida uruguay
236 concepcion del uruguay
```

```
==> venezuela.txt <==
282209 venezuela
69371 caracas venezuela
17085 valencia venezuela
16934 maracaibo venezuela
7116 barquisimeto venezuela
6894 maracay venezuela
6416 merida venezuela
4558 venezuela caracas
4242 mérida venezuela
3582 barinas venezuela
```