# Forecasting Protests by Detecting Future Time Mentions in News and Social Media

Sathappan Muthiah

Virginia Tech

July 2nd, 2014

# Table of Contents

# Problem Overview

- Detecting Future time mentions in relevant media to build a protest forecasting system.
- Investigate the selective superiorities of Different Social Media.

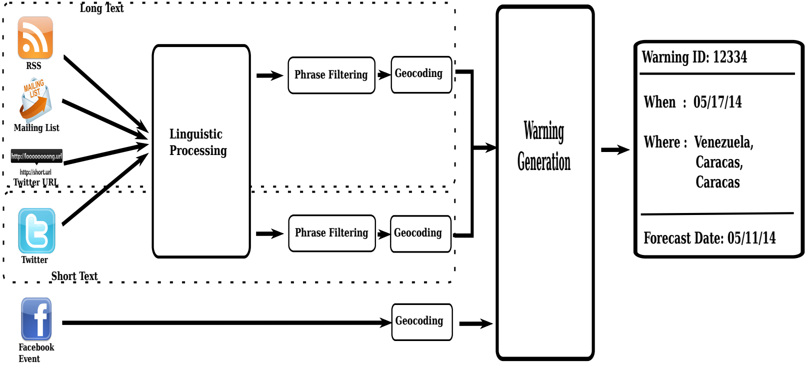
# Motivation

- Around 75% of the protests are planned, organized, or announced in advance.
- Identifying these planned protests is an easy way to forecast protests.

# Key Contributions

- Real-Time Prospective Study-most studies until now have been retrospective.
- Semi-Automatic approach for Learning Keyphrase filters.
- Handling Mutliple Sources.
- Reasoning about locations.
- Handling relative dates - some recent work use only absolute dates.

# Overall Framework



# Data Sources

- Long Text
  - RSS Feeds
    - News
    - Blogs
  - Twitter-URL
- Short Text
  - Twitter
- Facebook-Event

# Long Text - RSS Feeds

- Data Duration: November 2012 to March 2014
- 6540 News
- 6540 Blogs
- Talkwalker Alerts or Google Alerts

# Example - RSS Feed

# Short Text - Twitter

- Datasift Firehose
- Duration November 2012 to March 2014

# Example

# Facebook

- Facebook Graph API
- Facebook Query Language

# Example

# Preliminaries-Probabilistic Soft Logic

- Framework for collective probabilistic reasoning
- User defined predicates and rules
- MPE Inference

# Natural Language Enrichment

- Tokenization
- Lemmatization
- Noun Phrase Extraction
- Named Entity Extraction and Classification

# TIMEN Enrichment

- Extraction of Absolute Dates from text
- Identification of Relative dates like ‘yesterday, next wednesday’ etc.

# Geocoding- RSS Feeds

**Que la calle no calle**
A pesar de que el Gobierno insiste en promulgar la paz la concentración de ayer terminó con gases lacrimógenos. La GN volvió a salirse con las suyas y haciendo usos de las ballenas reprimieron otra manifestación pacífica, sin embargo, los estudiantes no se dan por vencidos y anunciaron que marcharán el domingo

La concentración convocada por el movimiento estudiantil en Caracas no culminó pacíficamente. Aunque desde las 11 de la mañana hasta las 2 de la tarde todo transcurrió con normalidad, a eso de las 2:30 pm, cuando la mayoría de los que se encontraban en la avenida Venezuela de El Rosal se disponían a irse, otros decidieron trasladarse hasta la autopista Francisco Fajardo para trancarla.

Fue en ese momento cuando efectivos de la Guardia Nacional accionaron sus bombas lacrimógenas contra los manifestantes para impedir que realizaran la toma.

Después la arremetida, a través de su cuenta twitter Juan Requesens, presidente de la Federación de Centros de Estudiantes de la Universidad Central de Venezuela (FCU-UCV) criticó que se hable de paz y luego se utilicen acciones violentas por parte de las fuerzas de seguridad: "Hablan de paz y después que los estudiantes nos concentramos pacíficamente gritando Ni un muerto más, nos lanzan bombas lacrimógenas".

El alcalde de Baruta, Gerardo Blyde, consider{o que fue "excesiva" la represión de la GN hacia los manifestantes en Las Mercedes. Pasadas las 4 de la tarde la arremetida contra los jóvenes continuó, esta vez desde la Plaza Altamira en Chacao.

El próximo domingo los universitarios esperan mantener la actividad de calle. Es por ello que convocaron a una marcha en la capital, donde esperan congregar a ciudadanos de todos los sectores que saldrán desde distintos puntos a la Plaza Brión, en Chacaito.

En las próximas horas deben confirmar ruta. "No nos arrodillamos seguiremos exigiendo justicia, igualdad y paz. Luchamos con el pueblo por sus derechos" Escribió Requesens.

→

{"admin1": "Caracas",
"city": "Caracas",
"country": "Venezuela",
"confidence": 0.42186905915279704}

{"admin1": "Miranda",
"city": "Baruta",
, "country": "Venezuela",
"confidence": 0.2639358965025394}

{"admin1" : "Ciego de Ávila",
"city" : "Venezuela",
"country" : "Cuba",
"confidence": 0.05116227467273876}

{"admin1": "Miranda",
"city": "Chacao",
, "country": "Venezuela",
"confidence": 0.2639358610172565}

{"admin1": "Cundinamarca",
"city": "El Rosal",
, "country": "Colombia",
"confidence": 0.0011984789871345436}

→

Admin1 : Caracas
City : Caracas
Country : Venezuela
Confidence : 0.42186905915279704

# Geocoding- RSS Feeds Contd.

- Primary rules

$$ENTITY(L, location) \tilde{\wedge} REFERSTO(L, locID) \rightarrow PSLLOCATION(Article, locID)$$

$$ENTITY(C, location) \tilde{\wedge} IsCountry(C) \rightarrow ArticleCountry(Article, C)$$

$$ENTITY(S, location) \tilde{\wedge} IsState(S) \rightarrow ArticleCountry(Article, S)$$

# Geocoding- RSS Feeds Contd.

- Secondary rules

$$ENTITY(O, organization) \tilde{\wedge} REFERSTO(O, locID) \rightarrow PSLLOCATION(Article, locID)$$

$$ENTITY(O, organization) \tilde{\wedge} IsCountry(O) \rightarrow ArticleCountry(Article, O)$$

$$ENTITY(O, organization) \tilde{\wedge} IsState(O) \rightarrow ArticleCountry(Article, O)$$

# Geocoding- Twitter

- Geotag of the tweet
- Twitter "places" metadata
- Other text fields (user profile, tweet text)

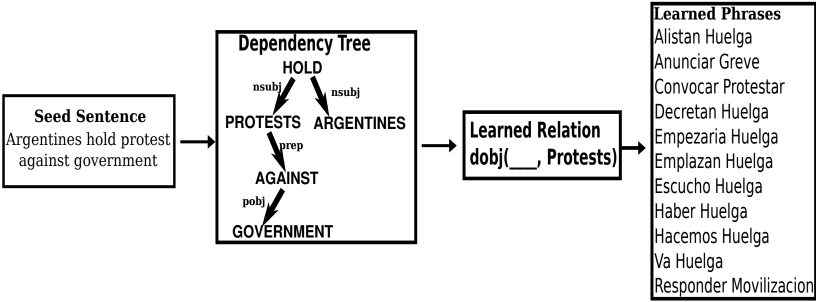
# Geocoding- Twitter

- Facebook Location Objects
- Facebook Event Venue/location

# Phrase List Development

- Semi-Automatic
- Different Lists are built for different Sources
- Seed phrases are identified from analysis of known planned events from print media.

# Dependency Parsing



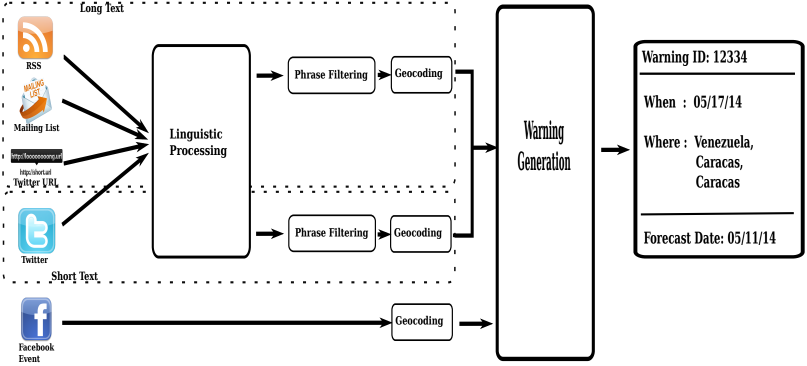
# Phrase List for Long Text

Example of phrases used for Long Text

# Phrase List for Short text

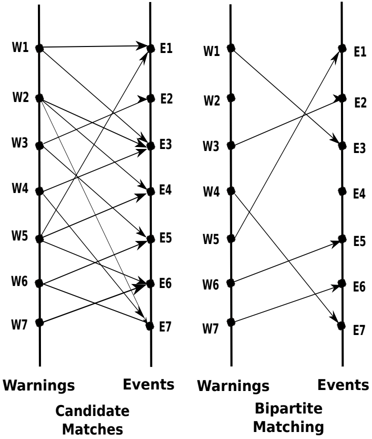
Example of Phrases used for Short text

# System Framework Once again



# Evaluation Methodology

- Bipartite Matching



# Evaluation Methodology Contd

- Date Score

$$\mathsf{LS} = 1 - \frac{\mathsf{min}(\text{distance offset}, 300)}{300} \tag{1}$$

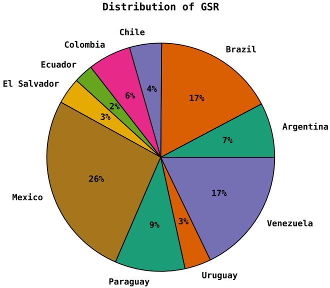
- Location Score

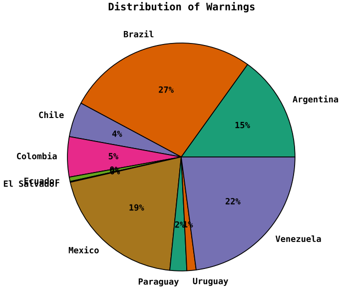
$$\mathsf{DS} = 1 - \frac{\mathsf{min}(\text{date offset}, \mathsf{INTERVAL})}{\mathsf{INTERVAL}} \tag{2}$$

- Total Quality Score

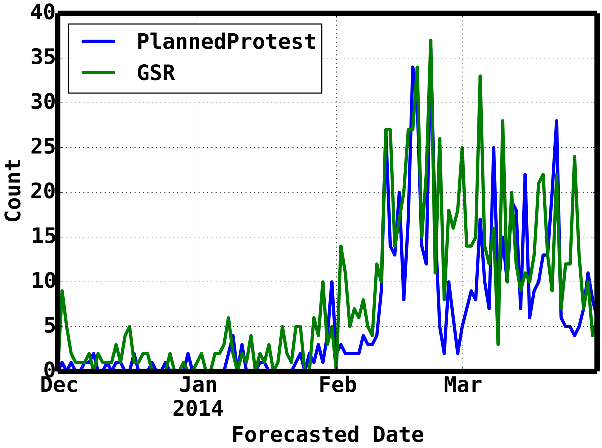
$$\mathsf{QS} = (\mathsf{DS} + \mathsf{QS}) * 2 \tag{3}$$

# Warnings vs GSR

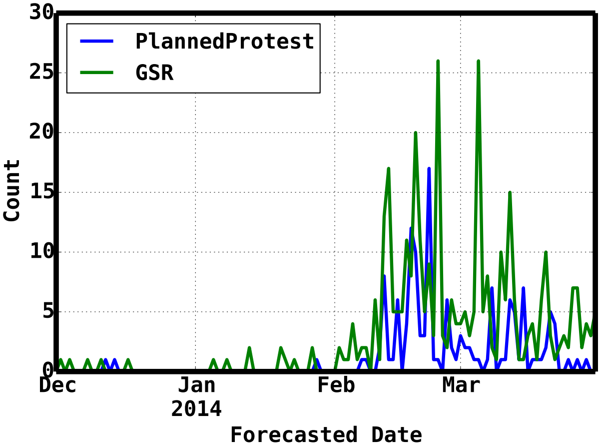


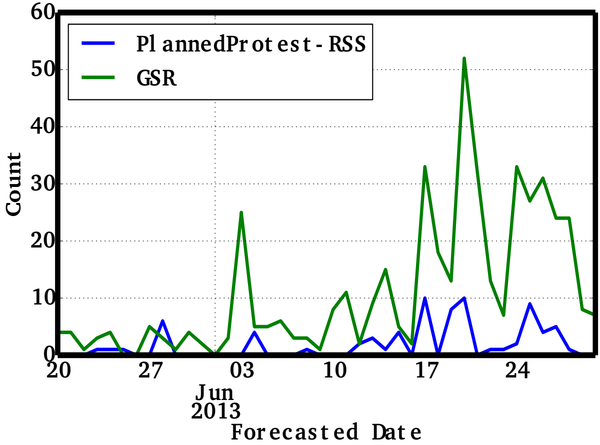
# Venezuelan Spring



# Venezuelan Violent Protests



# Brazilian Spring



# Individual Source Level Perfomance

| | News/Blogs | | | | Twitter | | | | Facebook | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | QS | Pr. | Rec. | LT | QS | Pr. | Rec. | LT | QS | Pr. | Rec. | LT | QS |
| AR | 3.14 | 0.32 | 0.69 | 3.94 | 3.52 | **0.78** | 0.14 | 3.14 | **3.70** | 0.50 | 0.04 | 3.00 | 3.02 |
| BR | 3.14 | 0.48 | 0.54 | **5.85** | - | - | - | - | **3.62** | **0.76** | 0.18 | 2.46 | 3.28 |
| CL | 3.06 | 0.91 | 0.67 | 5.40 | **3.52** | **1.00** | 0.23 | 4.29 | - | - | - | - | 3.16 |
| CO | 2.74 | 0.90 | 0.56 | **7.44** | 3.30 | **1.00** | 0.15 | 2.43 | **4.00** | **1.00** | 0.02 | 2.00 | 2.88 |
| EC | - | - | - | - | **2.32** | **1.00** | **0.06** | **17.00** | - | - | - | - | **2.32** |
| MX | 2.96 | 0.88 | 0.25 | **3.69** | 3.14 | **1.00** | 0.02 | 1.43 | **3.72** | 0.67 | 0.01 | 2.00 | 3.00 |
| SV | **3.22** | **1.00** | **0.03** | **1.0** | - | - | - | - | - | - | - | - | **3.22** |
| PY | 3.38 | **1.00** | **0.16** | 9.11 | 3.84 | **1.00** | 0.04 | **11.40** | 3.96 | **1.00** | 0.01 | 2.00 | 3.60 |
| UY | **3.24** | **1.00** | **0.29** | **2.40** | - | - | - | - | - | - | - | - | 3.24 |
| VE | **3.80** | **1.00** | 0.36 | **3.27** | 3.68 | 0.97 | 0.33 | 2.39 | - | - | - | - | 3.64 |
| ALL | 3.34 | 0.69 | 0.35 | **4.57** | 3.62 | **0.97** | 0.15 | 2.82 | 3.66 | 0.74 | 0.03 | 2.44 | 3.36 |

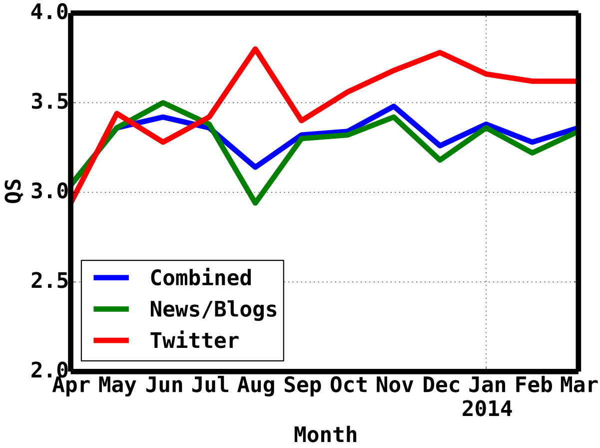
# RSS Feeds + Twitter-Urls

# Twitter

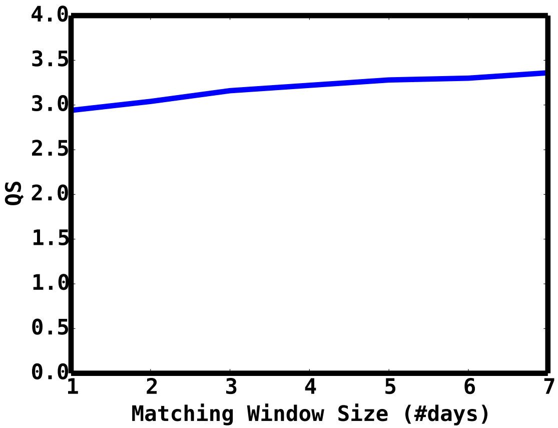
# Facebook

# all Sources

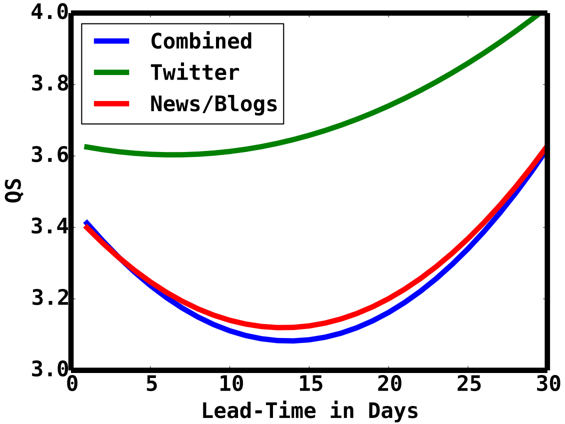
# Performance over time



# Quality Score vs Matching window size



# Lead-Time vs Quality



# Quality Score Distribution