

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

Sathappan Muthiah

Virginia Tech

July 2nd, 2014

Table of Contents

- 1 Problem Overview
 - motivation
- 2 Data Sources
- 3 Preliminaries
- 4 Linguistic Preprocessing
 - Natural Language Enrichment
 - TIMEN Enrichment
- 5 Geocoding
 - RSS
 - Twitter
 - Facebook
- 6 Phrase Filtering
- 7 Evaluation

Table of Contents

- 1 Problem Overview
 - motivation
- 2 Data Sources
- 3 Preliminaries
- 4 Linguistic Preprocessing
 - Natural Language Enrichment
 - TIMEN Enrichment
- 5 Geocoding
 - RSS
 - Twitter
 - Facebook
- 6 Phrase Filtering
- 7 Evaluation

Problem Overview

- Detecting Future time mentions in relevant media to build a protest forecasting system
- Investigate the selective superiorities of different social media

- 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 multiple sources
- Reasoning about locations
- Handling relative dates - some recent work use only absolute dates

Overall Framework

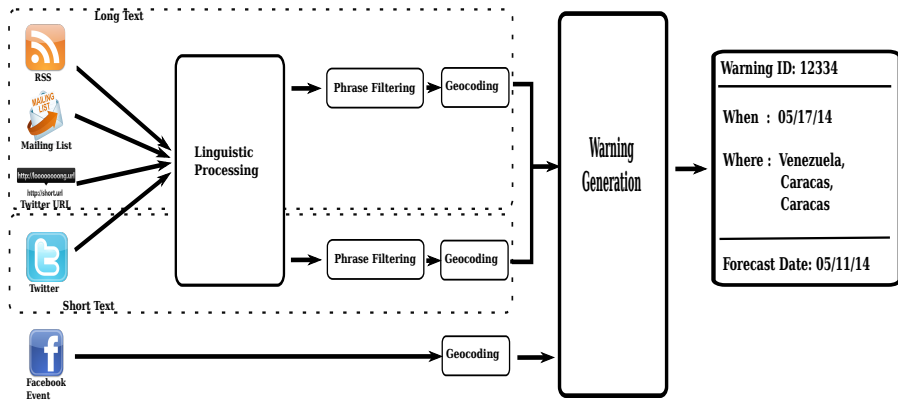


Table of Contents

- 1 Problem Overview
 - motivation
- 2 Data Sources
- 3 Preliminaries
- 4 Linguistic Preprocessing
 - Natural Language Enrichment
 - TIMEN Enrichment
- 5 Geocoding
 - RSS
 - Twitter
 - Facebook
- 6 Phrase Filtering
- 7 Evaluation

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

Long Text - RSS Feeds

- A total of 9498 different RSS feeds are ingest - 6236 news sources and 3262 blogs
- Duration: November 2012 to present
- List of news sources to ingest were obtained from Wikipedia, www.onlinenewspapers.com, LANIC, etc
- List of blogs were obtained from blog search engines like www.technorati.com
- Google/Talkwalker Alerts - Alerts for phrases in our keyphrase dictionary

Long Text - Example

TalCualDigital.com
CARACAS, miércoles 11 de junio, 2014

Inicio | Sobre mí | La Nación | Mundo | Economía | Opinión | Deportes | Ciudad | Entretenimiento

Que la calle no calle



Destacadas

Economía
Necesitan col
El presidente de Asef Marco, asegura que acordará las tarifas, en que la situación en el

Noticias
We are moving
Consejeros de la Asamblea Nacional, tras el anuncio de la Asamblea Nacional, se han trasladado a la sede de la

Noticias
Allow the UN!
El gobierno venezolano rechaza la propuesta de la Asamblea Nacional de permitir el ingreso de la ONU al país.

Noticias
They need to change more
El gobierno venezolano rechaza la propuesta de la Asamblea Nacional de permitir el ingreso de la ONU al país.

Noticias
What's there in jail
El gobierno venezolano rechaza la propuesta de la Asamblea Nacional de permitir el ingreso de la ONU al país.

La concentración convocada por el movimiento estudiantil en Caracas no cubrió 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 de Miranda para trancalear. 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 Riquelme, presidente de la Federación de Centros de Estudiantes de la Universidad Central de Venezuela (FCU-UCV), criticó que se había de paz y luego se usaron acciones violentas por parte de las Fuerzas de seguridad. "Habían de paz y después que los estudiantes nos concentráramos pacíficamente gritando ¡Ni un muerto más, nos lanzan bombas lacrimógenas!"

```
author: u'Daniele Persegani*',
'author_detail': {'name': 'Daniele Persegani*'},
'authors': {},
'content': 'STJ: receitas decorrentes da venda de im\x3f3veis comp\x3f5em PIS e Cofin .....',
'date': '2014-03-31T09:42:43',
'emblemsId': '4b0f1d0950fe15d89e9930822a1d9f677c36340f',
'emblemsLang': 'und',
'feed': 'rss-content-enriched',
'feedPath': [['rss-entries'], 'rss-content'],
'guidislink': False,
'id': 'http://www.jb.com.br/sociedade-aberta/noticias/2014/03/31/stj-receitas-.....',
'link': 'http://www.jb.com.br/sociedade-aberta/noticias/2014/03/31/stj-receitasca.....',
'links': [{'href': 'uhttp://www.jb.com.br/sociedade-aberta/noticias/2014/03/31/stj.....',
'rel': 'alternate',
'type': 'text/html'}],
'parentId': 'dae1d03d5407f1ac0e16ae5c7cc21d10357ac875',
'published': 'Mon, 31 Mar 2014 06:42:43 -0300',
'tags': [{'label': 'None', 'scheme': 'None', 'term': 'Sociedade Aberta'}],
'title': 'STJ: receitas decorrentes da venda de im\x3f3veis comp\x3f5em PIS e Cofins',
'title_detail': {'base': 'http://www.jb.com.br/sociedade-aberta/noticias/rss.xml',
'language': 'None',
'type': 'text/plain',
'value': 'STJ: receitas decorrentes da venda de im\x3f3veis comp\x3f5em..'},
'url': 'http://www.jb.com.br/sociedade-aberta/noticias/2014/03/31/stj-receitas-decor...',
'url_location': {'city': '', 'country': 'Brazil', 'state': ''}}
```

- Datasift Firehose
- Duration: November 2012 to present
- URL's mentioned in a tweet are fetched and used as a separate source (alongwith RSS feeds).

- Facebook Graph API - Query for Facebook Events that contain a particular keyword
- Facebook Query Language (FQL) - Obtain extra information of an Event-Id obtained by searching through Graph API



Table of Contents

- 1 Problem Overview
 - motivation
- 2 Data Sources
- 3 Preliminaries
- 4 Linguistic Preprocessing
 - Natural Language Enrichment
 - TIMEN Enrichment
- 5 Geocoding
 - RSS
 - Twitter
 - Facebook
- 6 Phrase Filtering
- 7 Evaluation

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

Table of Contents

- 1 Problem Overview
 - motivation
- 2 Data Sources
- 3 Preliminaries
- 4 Linguistic Preprocessing
 - Natural Language Enrichment
 - TIMEN Enrichment
- 5 Geocoding
 - RSS
 - Twitter
 - Facebook
- 6 Phrase Filtering
- 7 Evaluation

Natural Language Enrichment

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



Basis RLP

```
"BasisEnrichment":
{
  "language": "Spanish",
  "errorMessage": "",
  "enrichmentProcess": "RLP Java API v7.9.100"

  "nounPhrases": [{"expr": "Convocatoria", "offset":
"0:1"},
    { "expr": "Tercer Marcha de la
Dignidad Nacional http", "offset": "2:9"},
    .....
    "tokens": [{"lemma": "convocatoria", "POS":
"NOUNSG", "value": "Convocatoria"},
    {"lemma": ":", "POS": "PUNCT",
"value": ":"},
    .....
    "entities": [{"expr": "Dignidad Nacional",
"neType": "ORGANIZATION", "offset": "6:8"},
    { "expr": "http://t.co/qsCggp8Zv4",
"neType": "IDENTIFIER:URL", "offset": "8:15"},
    .....
    { "expr": "http://t.co/rA19epu3NX",
"neType": "IDENTIFIER:URL", "offset": "15:22"}
    ],
    "langSource": "Basis-RLP"
}
```

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



Basis RLP

```
"eventSemantics":
{
  "datetimes":
  [
    {
      "date": "2014-05-09",
      "phrase": "Miércoles",
      "offset": "9:10",
      "type": "implicit"
    },
    {
      ...
    }
  ]
}
```

Table of Contents

- 1 Problem Overview
 - motivation
- 2 Data Sources
- 3 Preliminaries
- 4 Linguistic Preprocessing
 - Natural Language Enrichment
 - TIMEN Enrichment
- 5 **Geocoding**
 - RSS
 - Twitter
 - Facebook
- 6 Phrase Filtering
- 7 Evaluation

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** y **el Rosa** 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. "Hablen 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**, considero 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 **Cinza Altamira** en **Caracas**.

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

En las próximas horas deben confirmarse 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

- Primary rules

$$\begin{aligned} ENTITY(L, location) \tilde{\wedge} REFERSTO(L, locID) \\ \rightarrow PSLLOCATION(Article, locID) \end{aligned}$$
$$\begin{aligned} ENTITY(C, location) \tilde{\wedge} IsCountry(C) \\ \rightarrow ArticleCountry(Article, C) \end{aligned}$$
$$\begin{aligned} ENTITY(S, location) \tilde{\wedge} IsState(S) \\ \rightarrow ArticleCountry(Article, S) \end{aligned}$$

- Secondary rules

$$\begin{aligned} &ENTITY(O, organization) \tilde{\wedge} REFERSTO(O, locID) \\ &\rightarrow PSLLOCATION(Article, locID) \end{aligned}$$
$$\begin{aligned} &ENTITY(O, organization) \tilde{\wedge} IsCountry(O) \\ &\rightarrow ArticleCountry(Article, O) \end{aligned}$$
$$\begin{aligned} &ENTITY(O, organization) \tilde{\wedge} IsState(O) \\ &\rightarrow ArticleCountry(Article, O) \end{aligned}$$

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

- Facebook Locations - similar to twitter places
- Facebook Event Venue tag
- Nearest geocoded point search using KD-Tree algorithm

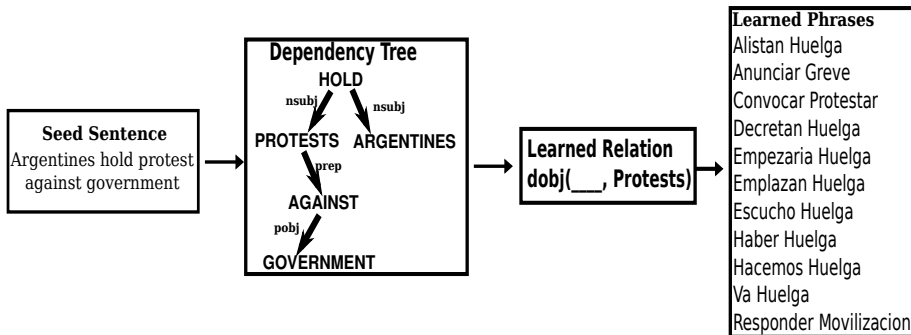
Table of Contents

- 1 Problem Overview
 - motivation
- 2 Data Sources
- 3 Preliminaries
- 4 Linguistic Preprocessing
 - Natural Language Enrichment
 - TIMEN Enrichment
- 5 Geocoding
 - RSS
 - Twitter
 - Facebook
- 6 Phrase Filtering**
- 7 Evaluation

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

ANUNCIAR GREVE
COMEÇAR GREVE
OUVIR GREVE
TEM GREVE
GREVE IR PASSAR
PREPARAÇÃO DE GREVE
ANUNCIAR MOBILIZAÇÃO
ANUNCIAR MOBILIZAÇÃO
PRÓXIMO MOBILIZAÇÃO
FAZER MOBILIZAÇÃO
MANHÃ DE MOBILIZAÇÃO
RADICALIZAR PROTESTAR
IR PARA ESSE MOBILIZAÇÃO
FAZER MOBILIZAÇÃO
VOSOTROS MANIFESTANTE SE REUNIRAM
CAMINHAR POR O RUA
ACOMPANHAR O PROTESTO
PROTESTARAM CONTRA
PROTESTO EM O BRASIL

ORGANIZE DEMONSTRATION
ORGANIZE PROTEST
ORGANIZE STRIKE
ORGANIZE WORK STOPPAGE
PLAN MARCH
PLAN MARCH
PLAN PROTEST
PLAN STRIKE

CONVOCAN HUELGA
DECRETAN HUELGA
EMPEZARIA HUELGA
PREPARANDO HUELGA
ALISTAN MOVILIZACIONES
INVITANDO MOVLIZACION
PROXIMO MOVLIZACION
MOVILIZACION MANANA
MOVILIZACION SABER
VAN A HUELGA
CONVOCAR PROTESTA
CONVOCAR HUELGA

Phrase List for Short text

MARCHAR	ANNOUNCE MOBILIZATION
GO PROTEST	JOIN MOVEMENT
PROTESTA	MOVILIZACIÓN
MARCHARAN	UNETE PROTESTA
RESPONDER MOVILIZACION	VENIR PROTESTA
MOVILIZACION MANANA	PROTESTAR
MOVILIZACION SABER	VAMOS PROTESTAR
ANUNCIAR PARO	PROTESTANDO CITO
CONFIRMAN PARO	MARCHAN CITO
INFILTREN PARO	PROTESTAR VER
HUELGA	INVITANDO MOVLIZACION
HABER HUELGA	PARTICIPAR MOVILIZACION
HACEMOS HUELGA	PROXIMO MOVLIZACION
VA HUELGA	REALIZAR MOVILIZACION
	RESPONDER MOVILIZACION

System Framework Once again

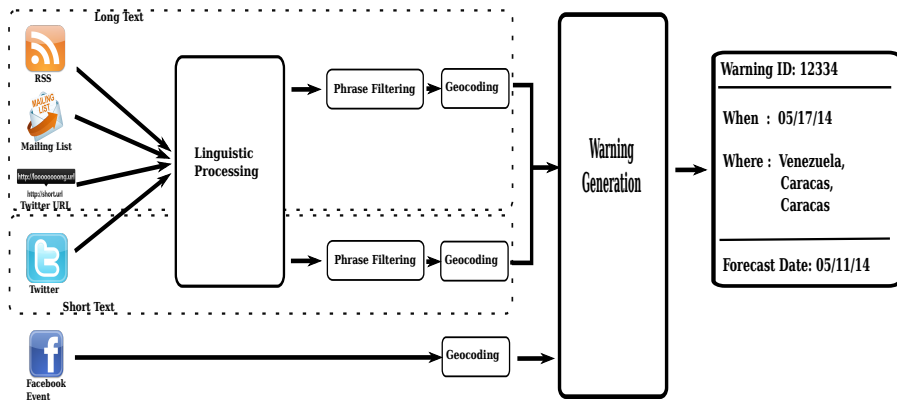
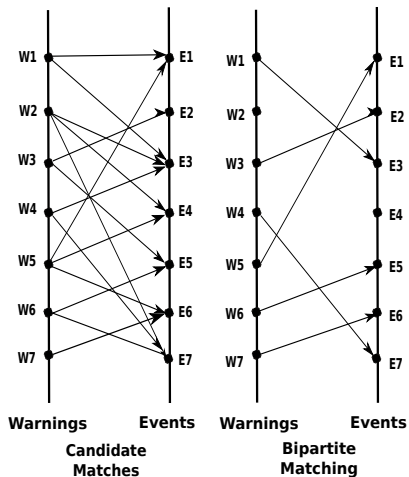


Table of Contents

- 1 Problem Overview
 - motivation
- 2 Data Sources
- 3 Preliminaries
- 4 Linguistic Preprocessing
 - Natural Language Enrichment
 - TIMEN Enrichment
- 5 Geocoding
 - RSS
 - Twitter
 - Facebook
- 6 Phrase Filtering
- 7 Evaluation

Evaluation Methodology

• Bipartite Matching



Evaluation Methodology (contd ...)

- Date Score

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

- Location Score

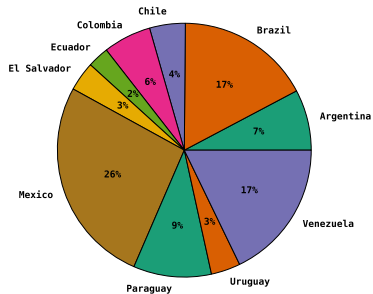
$$DS = 1 - \frac{\min(\text{date offset}, \text{INTERVAL})}{\text{INTERVAL}}$$

- Total Quality Score

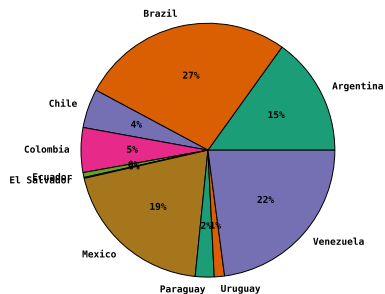
$$QS = (DS + LS) * 2$$

Warnings vs GSR

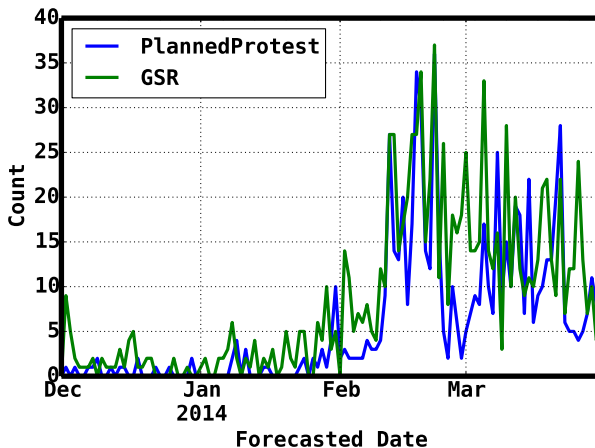
Distribution of GSR



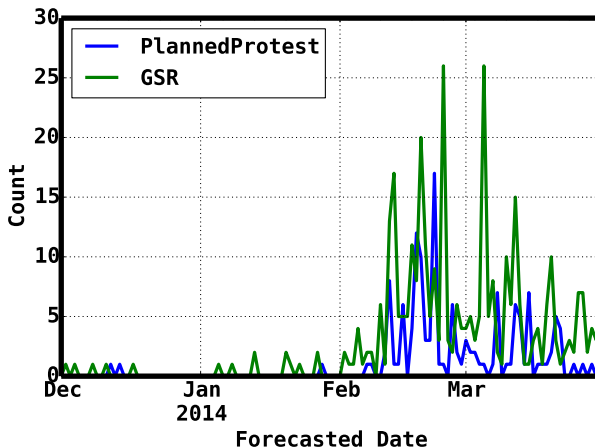
Distribution of Warnings



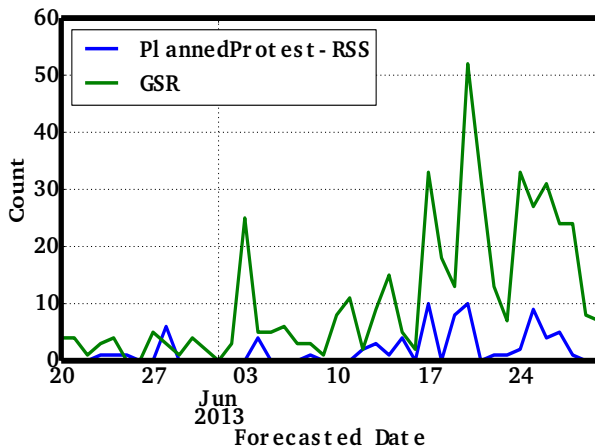
Venezuelan Spring



Venezuelan Violent Protests



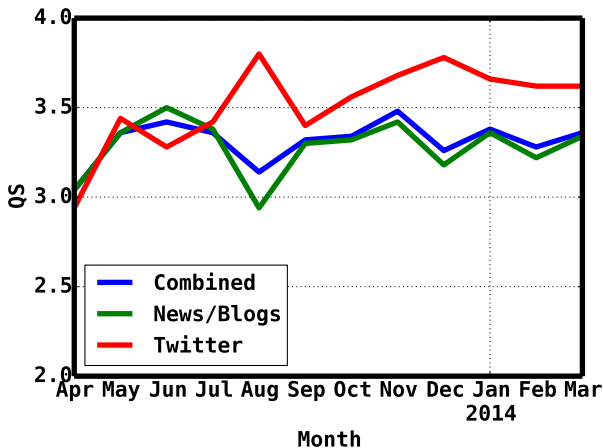
Brazilian Spring



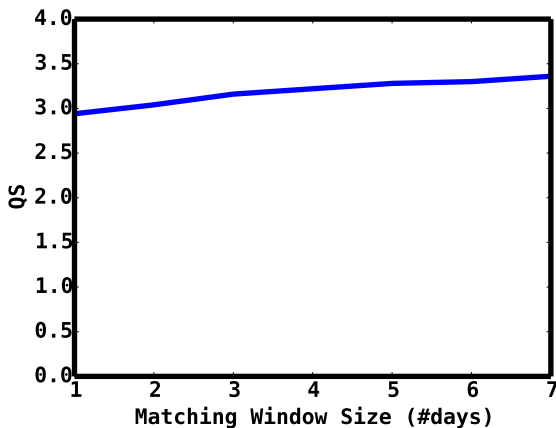
Individual Source Level Performance

	News/Blogs				Twitter				Facebook				Combined			
	QS	Pr.	Rec.	LT	QS	Pr.	Rec.	LT	QS	Pr.	Rec.	LT	QS	Pr.	Rec.	LT
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	0.36	0.80	4.50
BR	3.14	0.48	0.54	5.85	-	-	-	-	3.62	0.76	0.18	2.46	3.28	0.49	0.65	5.15
CL	3.06	0.91	0.67	5.40	3.52	1.00	0.23	4.29	-	-	-	-	3.16	0.83	0.80	5.92
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	0.84	0.65	6.47
EC	-	-	-	-	2.32	1.00	0.06	17.00	-	-	-	-	2.32	0.50	0.06	17.00
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	0.87	0.27	3.51
SV	3.22	1.00	0.03	1.0	-	-	-	-	-	-	-	-	3.22	1.0	0.03	1.0
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	0.96	0.20	9.35
UY	3.24	1.00	0.29	2.40	-	-	-	-	-	-	-	-	3.24	1.00	0.29	3.24
VE	3.80	1.00	0.36	3.27	3.68	0.97	0.33	2.39	-	-	-	-	3.64	0.99	0.69	2.88
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	0.73	0.51	4.08

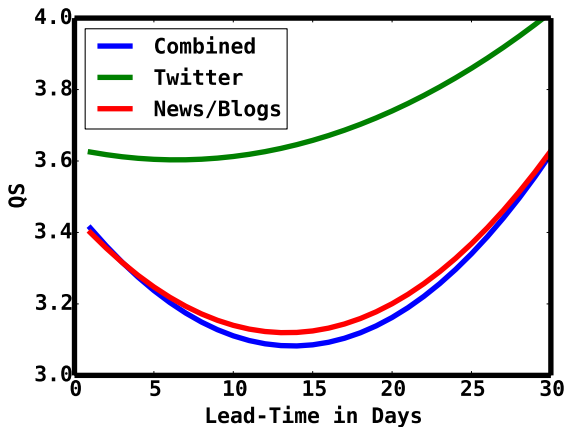
Performance over time



Quality Score vs Matching window size



Lead-Time vs Quality



Quality Score Distribution

