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

ABSTRACT

Civil unrest (protests, strikes, and ”occupy” events) is a common occurrence in both democracies and authoritarian regimes. The study of civil unrest is a key topic for political scientists as it helps capture an important mechanism by which citizenry express themselves. In countries where civil unrest is lawful, qualitative analysis has revealed that more than 75% of the protests are planned, organized, and/or announced in advance; therefore detecting future time mentions in relevant news and social media is a simple way to develop a protest forecasting system. We develop such a system in this paper, using a combination of key phrase learning to identify what to look for, probabilistic soft logic to reason about location occurrences in extracted results, and time normalization to resolve future tense mentions. We illustrate the application of our system to 10 countries in Latin America, viz. Argentina, Brazil, Chile, Colombia, Ecuador, El Salvador, Mexico, Paraguay, Uruguay, and Venezuela. Results demonstrate our successes in capturing significant societal unrest in these countries with an average lead time of 4.08 days. We also study the selective superiorities of news media versus social media (Twitter, Facebook) and identify relevant tradeoffs.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Theory

Keywords

ACM proceedings, LaTeX, text tagging

1. INTRODUCTION

[Sathappan writes: *Dr Mares's text*] Protest is a means by which citizens communicate their views and preferences to those in authority. Representative government is based upon communication between governed and governors; ideally, the channels for communication are open, transparent, credible and efficient. Governments, nevertheless, find it difficult to know on any one issue and at any one time how their constituencies value the available options. Elections are retrospective indicators and rarely issue specific; polling taps into sentiment, but is not a good indicator of priorities or strength of feeling because of the low cost associated with responding. Events, on the other hand, indicate a willingness to bear some costs (organization, mobilization, identification) in support of an issue and thus reveal not only preferences but provide some indication of priorities.

Protest is especially important in democracies that are struggling to consolidate themselves, such as those in Latin America. The combination of weak channels of communication between citizen and government, and a citizenry that still has not grasped the desirability of elections as the means to affect politics means that public protest will be an especially attractive option. To illustrate the power of protest in Latin America we need only recall that between 1985 and 2011 17 Presidents resigned or were impeached under pressure from demonstrations, usually violent, in the streets. Protests have also resulted in the rollback of prices increases for public services, such as in Brazil in June 2013.

We can hypothesize the protests that are larger will be more disruptive and communicate support for its cause better than smaller protests. Mobilizing large numbers of people is more likely to occur if a protest is organized and the time and place publicized. Because protest is costly and more likely to succeed if it is large we should expect planned, rather than spontaneous, protests to be the norm. Indeed, in a sample of 288 events from our study selected for qualitative review of their antecedents, for 225 we located communications regarding the upcoming occurrence of the event and only 49 were classified as ‘spontaneous’ (we could not determine whether communications had or had not occurred in 14 cases).

GRAHAM WRITES: With this in mind, we have sought to develop computational techniques for detecting these overt direct indicators of upcoming planned civil unrest events automatically from openly available sources of information. Detecting indicators for civil unrest planning activities might have application to a wide range of governmental and civil activity, from the issuance of travel warnings to rapid emergency response capabilities. Our detection models are relatively simple, making use of shallow linguistic analysis combined with targeted deep semantic analysis. Despite this simplicity, the models are able to detect indicators of event planning with surprisingly high accuracy. *More here summarizing results*

2. RELATED WORKS

Three categories of related work – *Event Detection, Extraction of Planned Events and Event Forecasting* – are briefly discussed here.

First, Event Detection/Extraction from textual News has been studied extensively in literature. [1] [32][13] make use of document clustering techniques to identify events retrospectively or as the stories arrive.[10],[5], [23] talk about extraction patterns/templates to extract information from text. [24] shows it's possible to accurately extract a calendar of significant events from Twitter by training a tagger for recognizing event phrases.[26] captures tweet clusters of interest to identify late breaking News from twitter.In an altogether different application [25] observes tweets to enable detection of occurences of Earth Quakes promptly.

Second, some work has been done regarding extraction of planned/future mentions of events from Social Media. RecordedFuture[29] [Sathappan writes: *TODO: write how it differs from our application*] is an analytics company that performs real-time analysis of news and tweets to identify mentions of future events.[28] and [30] use classification and regression techniques to identify the time to an event referred to by a tweet.In [28] a tweet was classified into one of the several identified equal length time bins. [15] tries to provide a collective image of the future associated with an entity summarizing all future related information available.In [7] and [6] content about known planned events is identified from Social Media. Several work has also been done on temporal information extraction from text. [18] presents a search engine ChronoSeeker for searching future and past events.It makes use of an SVM Classifier to disambiguate between the various temporal expressions in a document. .[4] and [12] also try to extract future temporal references from text, with the latter using a classifier approach to differentiate between a planned event and a rumor.

Third, few work has been done in the area of Event Forecasting. In [22] the authors learn event sequences from a corpora spanning over 22 years and then use these sequences to say if an event of interest (disease outbreaks, deaths and riots etc) will occur sometime in the future.. In [17], the author makes use of data from RecordedFuture[29] to find if a significant protest event will occur in the subsequent three days using a random forest classifier.The author only focuses on prediction of significant events and also the forecast is limited to the next three days.[11] and [31] are two pieces of research that are very close to our line of work. Both papers follow similar methodologies but are based on different datasets–Twitter and Tumblr.In [11], a list of 335 keywords identified by experts is used to filter twitter stream and the filtered twitter streams are then searched for the presence of future dates in a naive manner by first searching for month names and then for a number less than 31. Such an approach, will not be capable of finding relative mentions of future dates like ”tomorrow”, ”next tuesday” etc. Any location mentioned in the tweet text is used as the event location. If there are no location mentions then the location is determined based on [16].[31] works on Tumblr and makes uses of a fewer set of keywords(59) to filter the Tumblr feed. The filtered feed is further refined by searching for mentions of around 1022 different location names.Finally, the documents are searched for a future date in the same way as in [11].

3. PROBLEM

The problem we try to solve here is the identification of 'Calls for Protest/Strike or any Civil Disobedience movements' in Social Media like News, Blogs, Twitter, Facebook, etc. and to predict the date of event and event location upto a city level resolution.An accurately identified 'call for protest' is then sent out to an alerting system. An alert is structured as shown in Fig. 8. It is a structured record containing When/Where/Why/Who of the protest and the current date or date at which a forecast is made. The 'when' is specified in granularities of days. The **where** provides a tiered description specifying the (country, state, city), e.g., (Honduras, Francisco Morazan, Tegucigalpa). The **why** (or event type) captures the main objective or reason for a civil unrest event, and is meant to come from 7 broad classes (e.g., 'Employment & Wages', 'Housing', 'Energy & Resources' etc.) each of which is further categorized into whether the event is forecast to be violent or not. Finally, the **who** (or population) denotes common categories of human populations used in event coding [27] such as Business, Ethnic, Legal (e.g. judges or lawyers), Education (e.g. teachers or students or parents of students), Religious (e.g. clergy), Medical (e.g., doctors or nurses), Media, Labor, Refugees/Displaced, Agricultural (e.g. farmers, or just General Population.

Concomitant with the definitions in the above section, a GSR event contains again the where/why/when/who of a protest that has actually occurred and a *reported date* (the date a newspaper reports the protest as having happened). See Fig. 8 (right). The GSR is organized by an independent third party (MITRE) and the authors of this study do not have any participation in this activity.

4. PROBABILISTIC SOFT LOGIC

In this section, we briefly describe Probabilistic Soft Logic(PSL) [19] which is used to geo-code news/blogs as given in 5.1.1.

PSL is a framework for collective probabilistic reasoning on relational domains.PSL models have been developed in various domains, including collective classification [9], ontology alignment [8], personalized medicine [3], opinion diffusion [2] , trust in social networks [14], and graph summarization [21].PSL represents the domain of interest as logical atoms. It uses first order logic rules to capture the dependency structure of the domain, based on which it builds a joint probabilistic model over all atoms.Instead of hard truth values of 0 (false) and 1 (true), PSL uses soft truth values relaxing the truth vlaues to the interval $[0, 1]$.The logical connectives are adapted accordingly. This makes it easy to incorporate similarity or distance functions.

User defined *predicates* are used to encode the relationships and attributes and *rules* capture the dependencies and constraints. Each rule's antecedent is a conjunction of atoms and its consequent is a dis-junction. The rules can also labeled with non negative weights which are used during the inference process. The set of predicates and weighted rules

Table 1: comparison of our approach with other future event detection methods

	Domain Specific	Multi-Source	Geo-Coding	Temporal Normalization	Feature 4
[18, 30]				✓	
[31]	✓		✓		
reference set 3		✓			
our method	✓	✓	✓	✓	✓

thus make up a PSL program where known truth values of ground atoms derived from observed data and unknown truth values for the remaining atoms are learned using the PSL inference.

Given a set of atoms $\ell = \{\ell_1, \ldots, \ell_n\}$, an interpretation defined as $I : \ell \to [0,1]^n$ is a mapping from atoms to soft truth values. PSL defines a probability distribution over all such interpretaions such that those that satisfy more ground rules are more probable. *Lukasiewicz t-norm* and its corresponding co-norm are used for defining relaxations of the logical AND and OR respectively to determine the degree to which a ground rule is satisfied. Given an interpretation I, PSL defines the formulas for the relaxation of the logical conjunction ($\wedge$), disjunction ($\vee$), and negation ($\neg$) as follows:

$$\begin{aligned} \ell_1 \tilde{\wedge} \ell_2 &= \max\{0, I(\ell_1) + I(\ell_2) - 1\}, \\ \ell_1 \tilde{\vee} \ell_2 &= \min\{I(\ell_1) + I(\ell_2), 1\}, \\ \tilde{\neg} l_1 &= 1 - I(\ell_1), \end{aligned}$$

The interpretation I determines whether the rules is satisfied, if not, the *distance to satisfaction.* A rule $r \equiv r_{body} \to r_{head}$ is satisfied if and only if the truth value of head is atleast that of the body. The rule's distance to satisfaction measures the degree to which this condition is violated.

$$d_r(I) = \max\{0, I(r_{body} - I(r_{head}\}$$

PSL then induces a probability distribution over possible interpretations I over the given set of ground atoms l in the domain. If R is the set of all ground rules that are instances of a rule from the system and uses only the atoms in I then, the probability density function f over I is defined as

$$f(I) = \frac{1}{Z} \exp[-\sum_{r \in R} \lambda_r (d_r(I))^p] \quad (1)$$

$$Z = \int_I \exp[-\sum_{r \in R} \lambda_r (d_r(I))^p] \quad (2)$$

where λ_r is the weight of the rule r, Z is the continuous version of the normalization constant used in discrete Markov random fields, and $p \in \{1, 2\}$ provides a choice between two different loss functions, linear and quadratic. The values of the atoms can be further restricted by providing linear equality and inequality constraints allowing one to encode functional constraints from the domain. PSL provides for two kinds of inferences (a)most probable explanation and (b)calculation of the marginal distributions. In the MPE inference given a partial interpretation with grounded atoms based on observed evidence, the PSL program infers the truth values for the unobserved atoms satisfying the most likely interpretation. In the second setting, given ground truth data for all atoms we can learn the weights for the rules in our PSL program.

Table 2: EMBERS system statistics

Archived data	12.4 TB
Archive size	ca. 3 billion messages
Data throughput	200-2000 messages/sec
Daily ingest	15 GB
System memory	50 GB
System core	16 vCPUs
System output	ca. 40 warnings/day

5. LINGUISTIC PREPROCESSING

As part of the general streaming architecture of the EMBERS system, all textual input (e.g., tweets, news articles, blog postings) is subjected to shallow linguistic processing prior to analysis. Our data set is multilingual, with Spanish, Portuguese and English predominating in our data set. Commercial tools [1] are used for language identification, tokenization, lematization and named entity extraction. The lemmatized

Applying BASIS technologies' Rosette Language Processing (RLP) tools, the language of the text is identified, the natural language content is tokenized and lemmatized and the named entities identified and classified. Date expressions are normalized and deindexed (using the TIMEN [20] package). Finally, messages are geocoded with a specification of the location (city, state, country), being talked about in the message. An example of this enrichment processing can be seen in Fig. 1.

We make use of different geocoding methodologies for geocoding news/blogs and twitter.

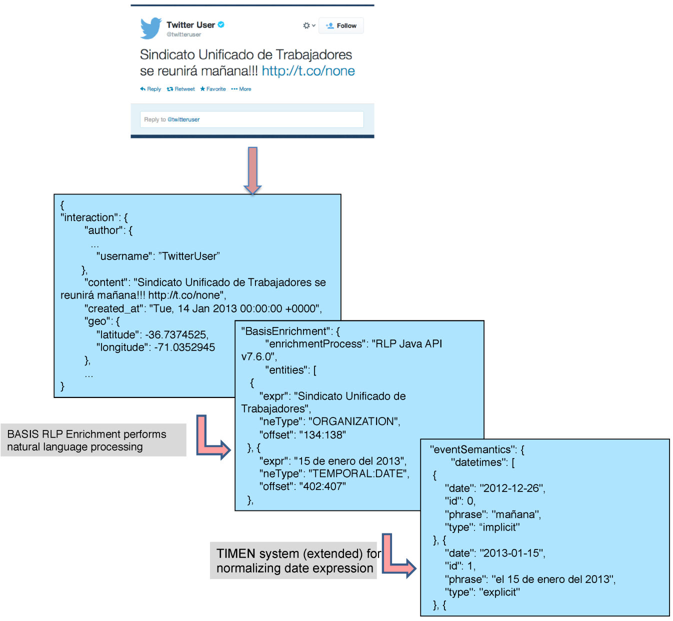
5.1 Geo-Coding

5.1.1 News/Blogs

To extract the protest location from news articles, we use *probabilistic soft logic* (PSL) described in 4 to build a model that performs robust, probabilistic inference given noisy signals. PSL takes a set of weighted, logic-like rules and converts them into a continuous probability distribution over the unknown truth values of logical facts. These truth values in PSL are relaxed into the $[0, 1]$ interval. We use this mechanism to build a model that infers the semantic location of an article by weighing evidence coming from the Basis entity extractions and information in the World Gazatteer.

The primary rules in the model encode the effect that Basis-extracted location strings that match to gazatteer aliases are indicators of the article's location, whether they be country, state, or city aliases. Each of these implications is conjuncted with an prior for ambiguous, overloaded aliases that is proportional to the population of the gazetteer location. For example, if the string "Los Angeles" appears in the ar-

[1] BASIS Technology's Rossette Linguistic Platform []



Twitter User
@twitteruser

Sindicato Unificado de Trabajadores se reunirá mañana!!! http://t.co/none

```
{
"interaction": {
        "author": {
          ...
            "username": ”TwitterUser”
        },
        "content": "Sindicato Unificado de Trabajadores se
reunirá mañana!!! http://t.co/none",
        "created_at": "Tue, 14 Jan 2013 00:00:00 +0000",
        "geo": {
            "latitude": -36.7374525,
            "longitude": -71.0352945
        },
        ...
}
```

```
"BasisEnrichment": {
        "enrichmentProcess": "RLP Java API
v7.6.0",
        "entities": [
  {
     "expr": "Sindicato Unificado de
Trabajadores",
     "neType": "ORGANIZATION",
     "offset": "134:138"
  }, {
     "expr": "15 de enero del 2013",
     "neType": "TEMPORAL:DATE",
     "offset": "402:407"
  },
```

```
"eventSemantics": {
      "datetimes": [
 {
    "date": "2012-12-26",
    "id": 0,
    "phrase": "mañana",
    "type": “implicit"
 }, {
    "date": "2013-01-15",
    "id": 1,
    "phrase": "el 15 de enero del 2013",
    "type": "explicit"
 }, {
```

BASIS RLP Enrichment performs natural language processing

TIMEN system (extended) for normalizing date expression

Figure 1: Message Enrichment

ticle, it could refer to either Los Angeles, California, or Los Ángeles in Argentina or Chile. Given no other information, our model would infer a higher truth value for the article referring to Los Angeles, California, because it has a much higher population than the other options.

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

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

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

The secondary rules, which are given half the weight of the primary rules, perform the same mapping of extracted strings to gazetteer aliases, but for extracted persons and organizations. Strings describing persons and organizations often include location clues (e.g., “mayor of Buenos Aires”), but intuition suggests the correlation between the article's location and these clues may be lower than with location strings.

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

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

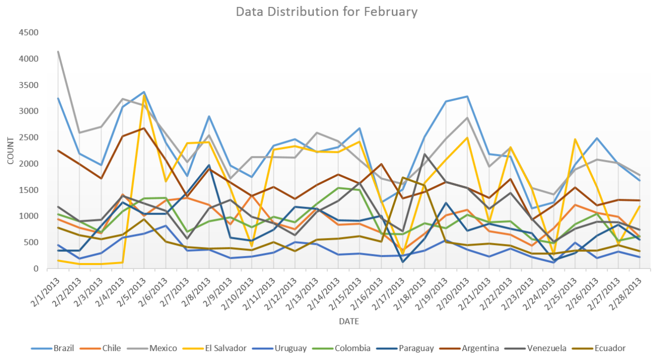


Figure 3: Rate of Arrival of News/Blogs

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

Finally, the model includes rules and constraints to require consistency between the different levels of geolocation, making the model place higher probability on states with its city contained in its state, which is contained in its country. As a post-processing step, we enforce this consistency explicitly by using the inferred city and its enclosing state and country, but adding these rules into the model makes the probabilistic inference prefer consistent predictions, enabling it to combine evidence at all levels.

$$PSLLOCATION(Article, locID) \tilde{\wedge} Country(locID, C) \\ \rightarrow ArticleCountry(Article, C)$$

$$PSLLOCATION(Article, locID) \tilde{\wedge} Admin1(locID, S) \\ \rightarrow ArticleState(Article, S)$$

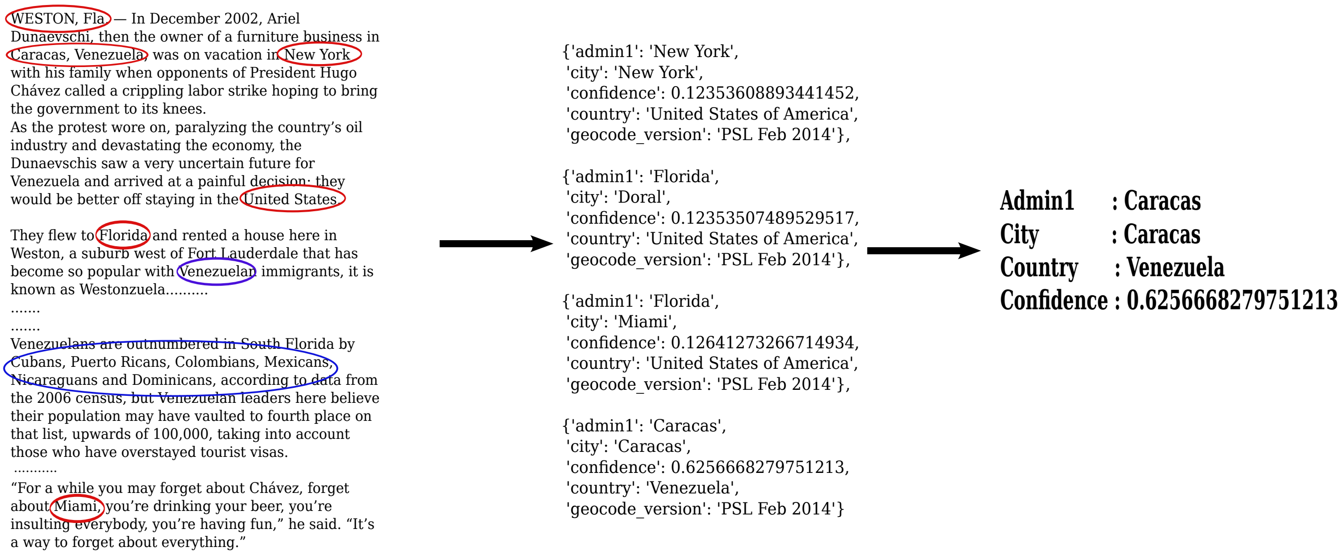
5.1.2 Twitter

The Twitter geocoding is achieved by first considering the most reliable but least available source, viz. geotags, which give us exact geographic locations that can be reverse geocoded into place names. We make use of a KD-Tree build using World Gazetteer data for this purpose.Second, we consider Twitter places and use place names present in these fields to geocode the place names into geographical coordinates. Finally, we consider the text fields contained in the user profile (location, description) as well as the tweet text itself to find mentions of relevant locations which can then be geocoded into geographical coordinates.

5.1.3 Facebook

We make use of only the Facebook Event data for our experiments. Facebook Events that have a venue are only used. A venue of a Facebook Event generally contains a latitude, longitude, country, state, city, street, etc. Under cases where only latitude and longitude is given we do a reverse-geocoding similar to what is used for Twitter to get the place names.

We use a highly specific list of 15 keywords and their variants in the three languages English, Spanish, Portuguese



WESTON, Fla. — In December 2002, Ariel Dunaevschi, then the owner of a furniture business in Caracas, Venezuela, was on vacation in New York with his family when opponents of President Hugo Chávez called a crippling labor strike hoping to bring the government to its knees.
As the protest wore on, paralyzing the country's oil industry and devastating the economy, the Dunaevschis saw a very uncertain future for Venezuela and arrived at a painful decision: they would be better off staying in the United States.

They flew to Florida and rented a house here in Weston, a suburb west of Fort Lauderdale that has become so popular with Venezuelan immigrants, it is known as Westonzuela..........
....
....
Venezuelans are outnumbered in South Florida by Cubans, Puerto Ricans, Colombians, Mexicans, Nicaraguans and Dominicans, according to data from the 2006 census, but Venezuelan leaders here believe their population may have vaulted to fourth place on that list, upwards of 100,000, taking into account those who have overstayed tourist visas.
....
"For a while you may forget about Chávez, forget about Miami, you're drinking your beer, you're insulting everybody, you're having fun," he said. "It's a way to forget about everything."

```
{'admin1': 'New York',
 'city': 'New York',
 'confidence': 0.12353608893441452,
 'country': 'United States of America',
 'geocode_version': 'PSL Feb 2014'},

{'admin1': 'Florida',
 'city': 'Doral',
 'confidence': 0.12353507489529517,
 'country': 'United States of America',
 'geocode_version': 'PSL Feb 2014'},

{'admin1': 'Florida',
 'city': 'Miami',
 'confidence': 0.12641273266714934,
 'country': 'United States of America',
 'geocode_version': 'PSL Feb 2014'},

{'admin1': 'Caracas',
 'city': 'Caracas',
 'confidence': 0.6256668279751213,
 'country': 'Venezuela',
 'geocode_version': 'PSL Feb 2014'}
```

```
Admin1     : Caracas
City       : Caracas
Country    : Venezuela
Confidence : 0.6256668279751213
```

Figure 2: Red circles denote named entities identified as locations and blue denotes other types of entities. The article is reported from Weston Florida US and talks about the recent increase of venezuelan population in the US compared to other Latin American Nations like Cuba etc.[Sathappan writes: *TODO: Replace with a better protest example*]

to query the Facebook Graph Api for Events.The Facebook Graph Api returns a list of Event-IDs, which is then used in an FQL query to get all available details of that event from Event table.

6. KEY PHRASE EXTRACTION

In Key Phrase Extraction, an input document is searched for the presence of one or more key phrases obtained in a semi-automatic manner as detailed in the following section.

Each key phrase mostly consists of atleast two words. The presence of a keyphrase is checked by searching for the presence of individual words of the keyphrase within the same sentence separated by utmost n words, where n is the average distance between the keyphrase words in the learning phase.

6.1 phraselearning

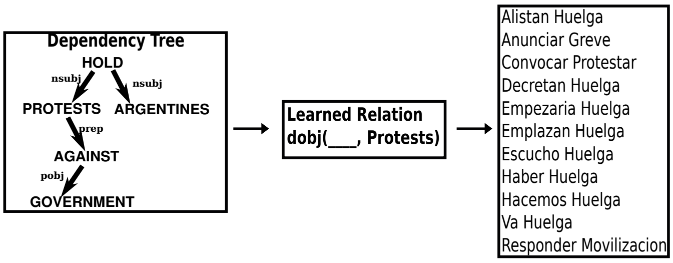
Different sets of phrases are used to extract events from News and Twitter.Usage of key phrases was found more accurate than using single keywords for extracting events of interest from the data stream.

Initially, a few seed phrases were obtained manually with the help of subject matter experts.

[Sathappan writes: *Jaime's Text –>*]

The idea was to have a simple query list for future civil unrest actions, not trying to identify the type of protest. The analysis of news reports for planned protests on the print media helped create a minimum set of words to use in the query. We choose four nouns from the basic query that is used predominantly to indicate a civil unrest in the print media - *Demonstration, March, Protest and Strike.* We translated them into Spanish and Portuguese, including synonyms. We then set some verbs that indicated a future action and included the proper future conjugations - to organize, to prepare, to plan, to announce, etc. For twitter, shorter terms were used to overcome its limitations of characters. These phrases had a more direct call for action, different from the ones chosen for RSS feeds. For example –*Marchar*, *manhã de mobilização*, *vamos protestar*, *Huelga* – was identified.

Figure 4: An Example of Phrase Learning



These phrases were then parsed using a dependency parser and the grammatical relationship between the core subject word—*protest*, *manifestación*, *Huelga*, etc.—and any accompanying word – *plan*, *call*, *anunciar* — was extracted [Sathappan writes: *TODO: Cite Freeling parser*].These grammatical relations serve as extraction patterns as in [23]. Extraction patterns identified in the last step is then used to learn more phrases from a corpora of filtered sentences extracted from the data stream of interest. Those sentences in a document that contained any one of the subject words and also had mentions of a future date was selected as part of the corpora on which the extraction patterns were used to learn more keyphrases.

The phrase learning is shown in Fig. 4

The set of learned phrases, is then cleansed by an expert to get the final set of key phrases. By this approach, we learned 112 phrases for News/blogs and 156 for tweets.

6.2 classification

For News/Blogs and Facebook, we make use of Text Based

Naive Bayes Classifier to identify the event-type and population. Unigram and Bi-gram word features are used for training the classifier.

For Twitter, as we send alerts based on a single tweet, we chose the event-type and population based on prior likelihood for that location.

7. APPROACH

All News,Blogs and Tweets are first filtered by searching for the presence of atleast one key phrase. The filtered documents are then searched for the mention of a future date. In case of News/blogs, the search for future date mentions is restricted to the sentence in which the keyphrase was found. where the phrase was found to reduce error. For tweets, no such restriction is made.

A warning/alert is then finally issued for those documents which also contain any location information. In the case of tweets, to avoid false alarms, we further filter the tweets by setting a threshold (set to 5) on the number of re-tweet of the tweet under consideration.

In the case of Facebook, a Facebook-Event is considered an alert if there are more number of attendees than number of rejects.

8. EXPERIMENTS

We evaluate the model under a strict and a relaxed condition. Under strict evaluation an alert can be matched to an event only if there is city level location match and the forecasted event date is same as the true event date. In the relaxed evaluation, we allow the matching to happen if the alert and the GSR event are within a 300 KM radius and the forecasted event date lies within a given interval of the true event date. We try different matching intervals ranging from 0 to 7. If x is the allowed interval, then a matching can happen if an alert's forecasted event date is within $+x$(upper bound) days of the GSR event and within a lower bound of $min(x, GSREventDate - DateofForecast)$

- **Perfomance over the months**

Fig. 7d provides the evaluation results of the model over the months with a source level breakdown. The QS reported is the weighted average of QS of all 10 countries where the weight for a country is the number of GSR events for that country. Twitter has a higher QS as multiple re-tweets of mention of future events in twitter is a direct indicator of the popularity of an event as well as the intent of people to join an event. While mention of Future events in News is simply a reporting of the event not much can be understood about the popularity of the event or about the people's support for the event.

- **Country-wise perfomance**

Table 3 presents the perfomance of the planned protest model (for March 2014) for each of the 10 countries of interest. It also presents a source wise breakdown.From the table it is evident that different data source prefer different countries. The News/Blogs data source produces alerts for most countries and also provides much higher recall as opposed to twitter which provides High QS but very little recall.Also News/Blogs has a higher lead-time of 4.57 days as compared to twitters 2.82.

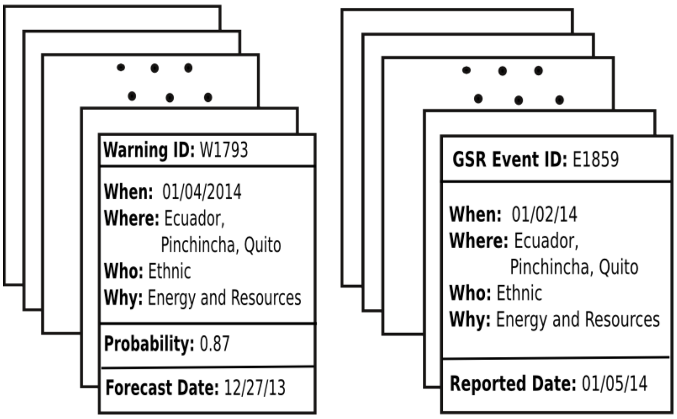


Figure 8: Structure of an Alert

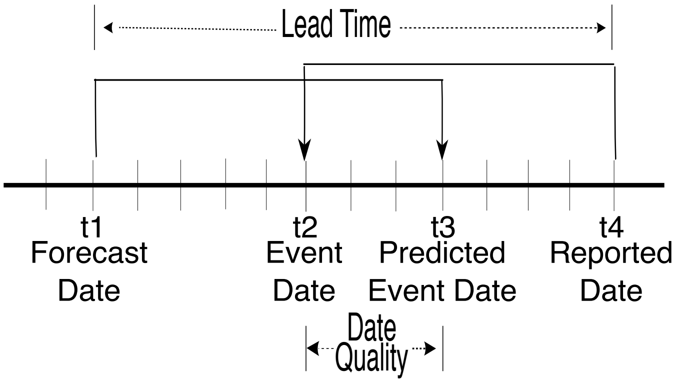


Figure 9: Matching Timeline

- **Case Study: Venezuelan and Brazilian Protests**

The recent Venezuelan protests against President Nicolas Maduro and the Brazilian Protests during June 2013 against bus fare hike were two significant protests during our period of evaluation. Fig. 7b7a show how well the planned protest model was able to predict the unfolding of events under both situations. Fig. 7c showcases the ability of the model to forecast the violent events also.

- **Lead-Time vs Quality Trade-Off**

Fig. 7e shows that the QS of the planned protest model increases with time.

- **Perfomance under stringent matching Criteria**

Fig. 7g shows the perfomance of the model when the matching window is varied from 7 to 1 in steps. We can see that the model is not affected badly even under the strict matching interval of 1-day difference.

- **Quality Score Distribution**

The hump on the right side of the Fig. 7f signifies that a majority of the planned protest alerts are high quality.

9. REFERENCES

[1] J. Allan, editor. *Topic Detection and Tracking: Event-based Information Organization.* Kluwer Academic Publishers, 2002.

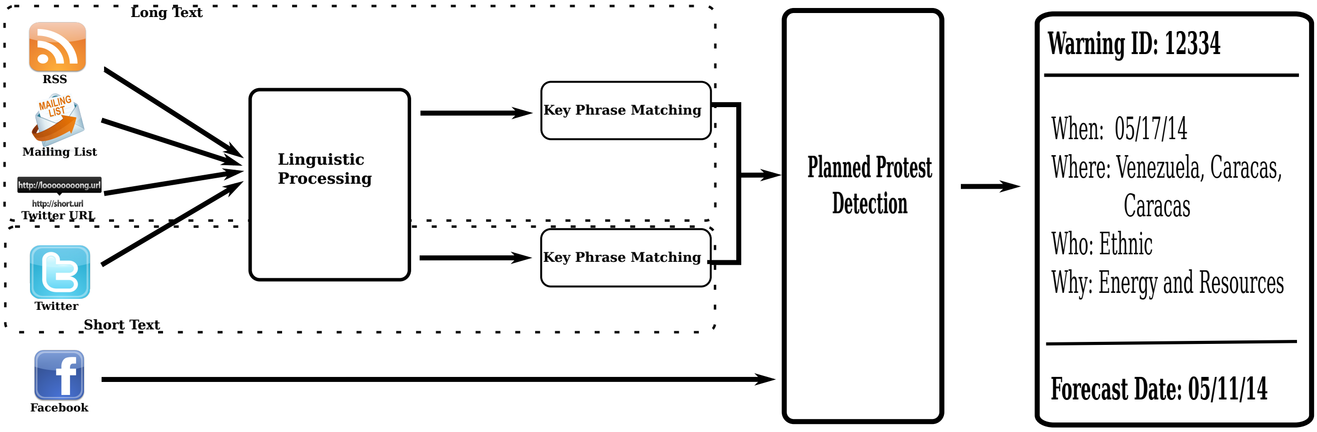
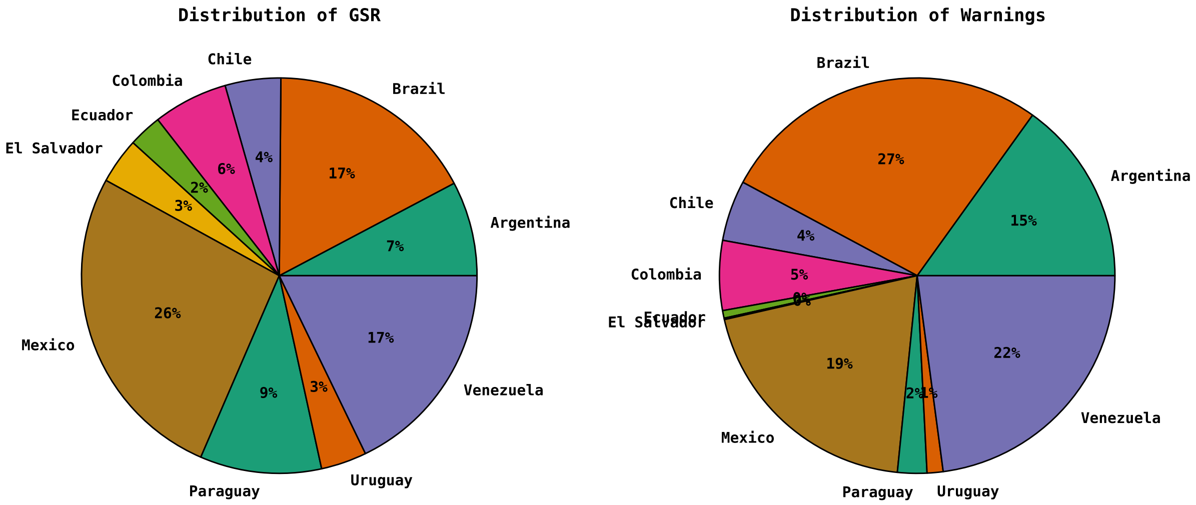


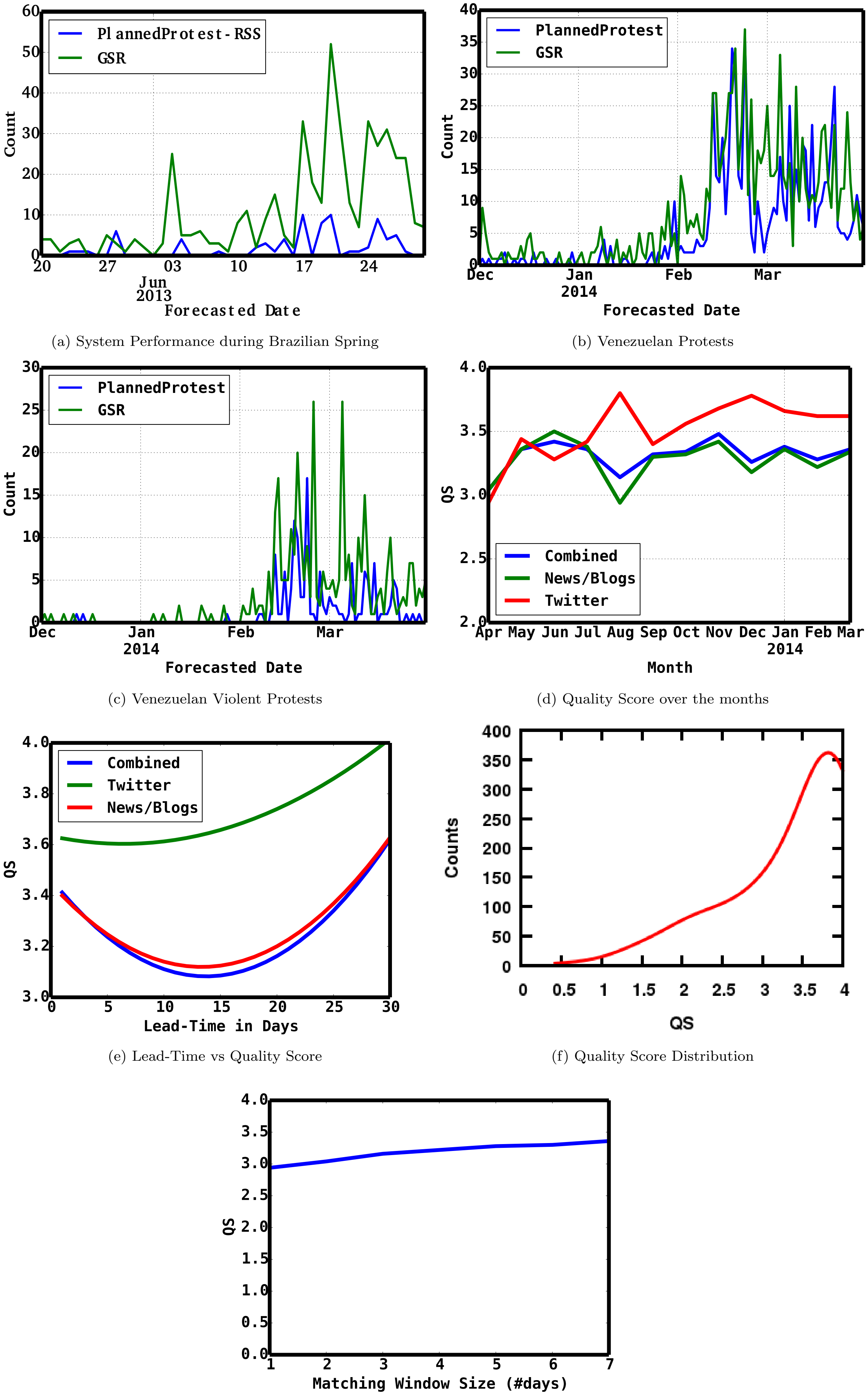
Figure 5: A diagram showing various steps of the Model



(a) GSR Distribution From 2012-11 to 2014-03

(b) Alerts Distribution From 2012-11 to 2014-03

Figure 6: Distribution of Alerts and GSR



(a) System Performance during Brazilian Spring

(b) Venezuelan Protests

(c) Venezuelan Violent Protests

(d) Quality Score over the months

(e) Lead-Time vs Quality Score

(f) Quality Score Distribution

(g) QS vs Matching Interval Trade-Off

Table 3: Comparing forecasting accuracy of RSS vs Twitter

	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	0.00	0.00	0.00	0.00	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	0.00	0.00	0.00	0.00	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	0.00	0.00	0.00	0.00	2.32	1.00	0.06	17.00	0.00	0.00	0.00	0.00	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	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	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	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	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	0.00	0.00	0.00	0.00	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

[2] S. Bach, M. Broecheler, L. Getoor, and D. O'leary. Scaling mpe inference for constrained continuous markov random fields with consensus optimization. In *Advances in Neural Information Processing Systems*, 2012.

[3] S. H. Bach, M. Broecheler, S. Kok, and L. Getoor. Decision-driven models with probabilistic soft logic. In *NIPS Workshop on Predictive Models in Personalized Medicine*, 2010.

[4] R. Baeza-Yates. Searching the future. In *SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*, 2005.

[5] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *International Joint Conferences on Artificial Intelligence*, IJCAI, 2007.

[6] H. Becker, F. Chen, D. Iter, M. Naaman, and L. Gravano. Automatic identification and presentation of twitter content for planned events. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, ICWSM, 2011.

[7] H. Becker, D. Iter, M. Naaman, and L. Gravano. Identifying content for planned events across social media sites. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM, 2012.

[8] M. Brocheler, L. Mihalkova, and L. Getoor. Probabilistic similarity logic. *arXiv preprint arXiv:1203.3469*, 2012.

[9] M. Broecheler and L. Getoor. Computing marginal distributions over continuous markov networks for statistical relational learning. In *Advances in Neural Information Processing Systems*, 2010.

[10] N. Chambers and D. Jurafsky. Template-based information extraction without the templates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT, 2011.

[11] R. Compton, C. Lee, T.-C. Lu, L. De Silva, and M. Macy. Detecting future social unrest in unprocessed twitter data:"emerging phenomena and big data". In *IEEE International Conference on Intelligence and Security Informatics*, ISI, 2013.

[12] G. Dias, R. Campos, and A. Jorge. Future retrieval: What does the future talk about. In *SIGIR Workshop on Enriching Information Retrieval*, 2011.

[13] E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: Providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the 13th International Conference on World Wide Web*, WWW, 2004.

[14] B. Huang, A. Kimmig, L. Getoor, and J. Golbeck. Probabilistic soft logic for trust analysis in social networks. In *International Workshop on Statistical Relational AI*, 2012.

[15] A. Jatowt and C.-m. Au Yeung. Extracting collective expectations about the future from large text collections. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM, 2011.

[16] D. Jurgens. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Seventh International AAAI Conference on Weblogs and Social Media*, ICWSM, 2013.

[17] N. Kallus. Predicting crowd behavior with big public data. In *Proceedings of the 23rd international conference on World wide web*, WWW, 2014.

[18] H. Kawai, A. Jatowt, K. Tanaka, K. Kunieda, and K. Yamada. Chronoseeker: Search engine for future and past events. In *Proceedings of the 4th International Conference on Uniquitous Information Management and Communication*, ICUIMC, 2010.

[19] A. Kimmig, S. Bach, M. Broecheler, B. Huang, and L. Getoor. A short introduction to probabilistic soft logic. In *Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications*, 2012.

[20] H. Llorens, L. Derczynski, R. J. Gaizauskas, and E. Saquete. TIMEN: An open temporal expression normalisation resource. In *Proceedings of Language Resources and evaluation*, LREC, 2012.

[21] A. Memory, A. Kimmig, S. Bach, L. Raschid, and L. Getoor. Graph summarization in annotated data using probabilistic soft logic. *status: accepted*, 2012.

[22] K. Radinsky and E. Horvitz. Mining the web to predict future events. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM, 2013.

[23] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the conference on Empirical methods in natural language processing*, EMNLP, 2003.

[24] A. Ritter, Mausam, O. Etzioni, and S. Clark. Open

domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD, 2012.

[25] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW, 2010.

[26] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: News in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS, 2009.

[27] P. A. Schrodt. Automated production of high-volume, near-real-time political event data. In *American Political Science Association meetings*, 2010.

[28] H. Tops, A. van den Bosch, and F. Kunneman. Predicting time-to-event from twitter messages. In *Proceedings of the 25th Benelux Conference on Artificial Intelligence*, BNAIC, 2013.

[29] S. Truvé. Big data for the future: Unlocking the predictive power of the web. *Recorded Future, Cambridge, MA, Tech. Rep*, 2011.

[30] A. H. urriyetoglu, F. Kunneman, and A. van den Bosch. Estimating the time between twitter messages and future events. In *Proceedings of the 13th Dutch-Belgian Workshop on Information Retrieval*, DIR, 2013.

[31] J. Xu, T.-C. Lu, R. Compton, and D. Allen. Civil unrest prediction: A tumblr-based exploration. In *Social Computing, Behavioral-Cultural Modeling and Prediction*. 2014.

[32] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.