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

Sathappan Muthiah

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Naren Ramakrishnan, Chair
Chang Tien Lu
Graham E Katz

June 26, 2014
Arlington, Virginia

Keywords: Textmining, Information Retrieval, Social Media

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

Sathappan Muthiah

(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) to identify relevant tradeoffs.

Dedication

To my wonderful Mom, Dad and Brothers

Acknowledgments

First and foremost I would like to thank my advisor Dr.Naren Ramakrishnan. He was not only instrumental in kindling my interest in machine learning and data mining but is also my mentor in many ways. I also thank Dr Graham E Katz for his constant guidance and help with Software Integration.

I thank my Parents and Brothers for their great support without which i wouldnt have been able to complete.

I would like to thank my fellow graduate students Rupinder Paul Khandpur, Aravindan Mahendiran, Wei Wang, Fang jin, Prithwish Chakraborty, Parang Saraf, Saurav Ghosh and Michael Shuffet for their help and support throughout my graduate studies.

I would like to thank Shahriar Hossain and Patrick Butler for their time and help on answering several of my queries.Also, i would like to thank Nathan Self for his help with visualizations and Jaime Arredondo for his help with building the keyphrase list.

I also like to thank all our collaborators at CACI inc., Andrew Doyle, Ilya Zavorin, Chris Ackerman, Jim Ford and Zunsik Lim for their help with Software Testing and Integration.

Finally I would also like to thank my roommates Rajesh Subbiah, Arvinth Chanthar Rathinama Saran Kumar and Chao Yang for making my stay at Blacksburg a memorable one.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Civil unrest (protests, strikes, and "occupy" events) is a common happening in both democracies and authoritarian regimes. Although we typically associate civil unrest with disruptions and instability, for a social scientist civil unrest reflects the democratic process by which citizenry communicate their views and preferences to those in authority. The advent of social media has afforded citizenry new mechanisms for organization and mobilization, and online news sources and social networking sites like Facebook and Twitter can provide a window into civil unrest happenings in a particular country.

Our basic hypothesis is that 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 announced in advance. 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 (more details later), for 225 we located communications regarding the upcoming occurrence of the event in media, and only 49 could be classified as spontaneous (we could not determine whether communications had or had not occurred in the remaining 14 cases).

We sought to develop a computational protest forecasting system by identifying and mining mentions of future planned events in (news and social) media. Why is this problem difficult? Consider the example article shown in Fig. 1.1, covered in greater detail in Fig. 4.6. Written in Spanish and published in Venezuela, it describes plans to protest on 'Sunday' and mentions at least five different locations – of which at least two are ambiguous and can either be in Colombia, Cuba or Venezuela. Significant reasoning is required to discern the correct protest location and to identify the intended date from the vague reference of 'Sunday.' Once the location and time of the future protest are inferred, a suitably generated alert has applications to a wide range of governmental and civil activity, from the issuance of travel warnings to rapid emergency response capabilities.

TalCualDigital.com

CARACAS, miércoles 11 de junio, 2014

Inicio | Editorial | La Nación | Mundo | Economía | Opinión | Deportes | Ciudad | Entretenimiento

Que la calle no calle

1.9k Share | 544 twittear | 0 Enviar | 1 Comentar

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".

Destacadas

Economía
Necesitan col
El presidente de ANP
Marco, asegura que
acordar las tarifas en
que la situación es ir

Economy
We are moving
Consecomercio President Mauricio Tancredi, insists that economic measures are reversed and finished flexible access to foreign exchange

The Nation
Allow the UN!
VP coordinator Carlos Vecchio requires the government to allow the entry of the UN, with the intention of giving solution to the crisis in the coun ...

Economy
They need to charge more
ANPAGE President Benigno Luis Marco, ensures that they are searching agree rates in parking lots because the situation is unsustainable

The Nation
Want them in jail
The prosecutor Luisa Ortega Díaz requesting the arrest of Pedro Burelli, Diego Arria and Ricardo Koesling as they have the assurance that they are outside the country and not ...

Lo último | + vistas | + comentadas

- Saving Lives
- BOB WELCH
- We are moving
- It has to be released
- Football though poorly paid
- Allow the UN!
- Family labor
- 10 THINGS TO BRAZIL 2014
- They need to charge more

Figure 1.1: An example article describing plans for a future protest (Venezuela, June 11, 2014). The red and black circle show the identified location and date.

Our detection approach combines shallow linguistic analysis to identify a corpus of relevant documents (articles, tweets) which are then subject to targeted deep semantic analysis. Despite its simplicity, we are able to detect indicators of event planning with surprisingly high accuracy. Our contributions are:

1. We present a protest forecasting system that couples three key technical ideas: key phrase learning to identify what to look for, probabilistic soft logic to reason about location occurrences in extracted results, and date normalization to resolve future tense mentions. We demonstrate how the integration of these ideas achieves objectives in precision, recall, and quality (accuracy).

2. 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. We conduct ablation studies to identify the relative contributions of news media (news + blogs) versus social media (Twitter, Facebook) to identify future happenings of civil unrest. Through these studies we illustrate the selective superiorities of different sources for specific countries.

3. Unlike many studies of retrospective forecasting of protests, we assess the lead time from when the forecast is made to the actual event date, to assess the forecasting prowess of our approach. Our results demonstrate that we are able to capture significant societal unrest in the above countries with an average lead time of 4.08 days. This illustrates that the approach here can be used in a practical protest forecasting system.

Chapter 2

Related Work

Five categories of related work are briefly discussed here.

2.1 Event detection via text extractions

Event Detection is an extensively studied topic in the literature. Document clustering techniques are used in [8, 9, 10] to identify events retrospectively or as the stories arrive. Works like [11, 12, 13] focus on extraction patterns (templates) to extract information from text. Ritter et al. [14] show that it is possible to accurately extract a calendar of significant events from Twitter by training a tagger for recognizing event phrases. Highly specialized applications also exist; e.g., Sakaki et al. [15] mine tweets to enable prompt detection of occurences of earthquakes.

2.2 Temporal information extraction

Temporal Information Extraction is another well studied topic. The TempEval challenge [16] led to a significant amount of algorithmic development for temporal NLP. For instance, a

Table 2.1: Comparison of our approach against other future retrieval techniques.

	Relative date resolution	Ingest multiple sources?	**Reasoning about location**	**Learning word/phrase filters**
'Future' Search Engines [1, 2, 3]	✓			
Time-to-Event Recognition [4, 5]	✓			
Planned Protest Detection [6, 7]		✓		
This thesis	✓	✓	✓	✓

specification language for temporal and event expressions in natural language text is described in [17]. Refs. [18] and [19] provide methods to resolve temporal expressions in text (our own work here uses the TIMEN package [18]).

2.3 Event forecasting

Event forecasting is a burgeoning area. Radinsky and Horvitz [20] find event sequences from a corpora and then use these sequences to determine if an event of interest (e.g., a disease outbreak, or a riot) will occur sometime in the future. This work predicts only if a potential event will happen given a historical event sequence but does not geolocate the event to a city-level resolution, as we do here. Kallus [21] makes use of event data from RecordedFuture [22] to determine if a significant protest event will occur in the subsequent three days and casts this as a classification problem. This work only focuses on prediction of significant events (suitably defined) and the forecast is limited to the next three days. Ramakrishnan et al. [23] describe the EMBERS system for forecasting civil unrest using open source indicators but this work is primarily focused on shallow mining of a broad set of data sourceas in contrast to the focused analysis of planned protest announcements that we study here.

2.4 Future Retrieval

Finally, Future Retrieval, an emerging research topic, is another area of research most closely related to our work. Baeza-Yates [3] providing one of the earliest discussions of this topic; here future temporal information in text is found and used to retrieve content from search queries that combine both text and time with a simple ranking scheme. Kawai et al. [1] present a search engine (ChronoSeeker) for searching future and past events. They make use of an SVM classifier to disambiguate between the various temporal expressions in a document. Dias et al. [24] classify web snippets into three classes depending on if a future date can/cannot be predicted from the snippet or if it is a rumor. RecordedFuture [22], introduced earlier, conducts real-time analysis of news and tweets to identify mentions of events along with associated times. Anectodally it is estimated that approximately (only) 5–7% of events extracted by RecordedFuture are about the future. Tops et al. [4] aim to classify a tweet talking about an event into discrete time segments and thereby predict the 'time to event'. Bosch et al. [5] use regression techniques to identify the time to an event referred to by a tweet. Jatowt et al. [2] provide a collective image of the future associated with an entity summarizing all future related information available. Becker et al. [25] try to identify more content about known planned events (e.g., a concert) across social media. This work for instance assumes that we know the event beforehand and aims to identify relevant details of the event.

In particular, two publications—Compton et al. [7] and Xu et al. [6]—align very closely to our own work as their emphasis is on protest forecasting. Both works are aimed at forecasting protests but emphasize different data sources and different methodologies. For instance, the work in [7] filters the Twitter stream for keywords of interest and searches for future date mentions in only absolute terms, i.e., explicit mentions of a month name and a number (date) less than 31. Such an approach will not be capable of extracting the more common way in which future dates are referenced, e.g., phrases like “tomorrow,” “next tuesday.” The work in [6] by the same group of authors uses the Tumblr feed with a smaller set of keywords but again is restricted to the use of absolute time identifiers.

In surveying the state-of-the-art, we arrived at desiderata for a planned protest forecasting system. As shown in Table 2.1, we desire a system that is capable of: i) ingesting a broad range of data sources from both popular news and social media, ii) learning relevant phrases for tracking protests, iii) handling relative mentions of dates, and iv) providing a rich representational and reasoning basis for location. As Table 2.1 summarizes, current systems provide only partial solutions and the proposed approach addresses all four desired criteria.

2.5 Geocoding

A News report, blog posting or a tweet can have multiple locations associated with it. It is necessary to disambiguate each of these locations and try to identify what they refer to – whether they refer to the location from where the document is written, users home location or the geo-focus of the content etc.

Identifying the different geographical locations associated with a tweet/news document is of great interest to the researchers recently. David et al. [26], Lindamood et al. [27] and Backstrom et al. [28] talk about using the network information based on social relationships to infer an users location.These approaches infer locations by spatially propagating location assignments through the social network, using a samll number of known users/locations. Such approaches are mostly based on the assumption that people we interact with on a daily basis always live near us.

Amitay et al. [29], Fink et al. [30],Cheng et al. [31], etc., use content based approaches to determine a documents location. Yin et al. [32] introduces a topic modelling approach called Latent Geographical Topic Analysis, that combines location and text to identify topics specific to a geographical area. These topics can then be used as important cures to group different geographical regions and also to identify the location of new users based on the content they publish.

Some work has also been done where both knowledge from the content and network are used to infer the location. In [33] Li et al. builds a unified discriminative influence model to combine both socail network information and the user-centric information available from his/her tweeting history using a probabilistic framework. Li et al., in yet another work [34], try to capture the location profile of an user from their followers network and tweet history and also profile the users location for each avaliable relationship in the network.

For our purpose, we try to distinguish between an users home location and the tweets content location. Thus, we mainly make use of content based approaches for identifying the geographical focus of a document. The approach used is described in detail in section 4.6.

Chapter 3

Preliminaries

Our emphasis in this work is on Latin America. Protest is an important topic of study in this region, as many countries here are democracies struggling to consolidate themselves. 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 price increases for public services, such as during the 'Brazilian Spring' of June 2013. Our goal is to identify calls for protests, strikes, or civil disobedience movements from news, blogs, Tweets, and Facebook pages, with a view toward predicting the **when** (date of the event) and **where**, i.e., event location, up to city-level resolution, e.g., the city of *Tegucigalpa* in the state of *Francisco Morazan* in the country of *Honduras*). We refer to our forecasts as alerts or warnings (see Fig. 3.1 (left)). In looking at the structure of the alert, it is important to distinguish between the forecast date (when the forecast is made) and the predicted event date (i.e., the **when** of the event).

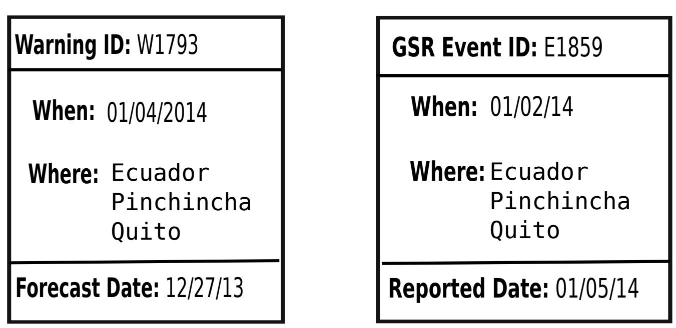


Figure 3.1: An example warning (left) and GSR (Goal Standard Report) event (right).

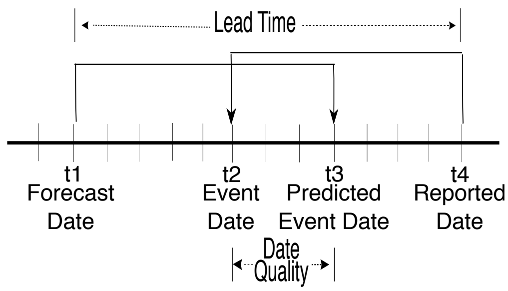


Figure 3.2: Alert sent at time $t1$ predicting an event at time $t3$ can be matched to a GSR event that happened at time $t2$ and reported at time $t4$ if $t1 < t4$.

3.1 Evaluation Metrics – Quality score

To evaluate our alerts we have access to a database of protests organized by a third party. We refer to this database as the GSR (for Gold Standard Report). Human analysts scan newspapers of record in the countries of interest and catalog protests. The structure of a GSR event (see Fig. 3.1 (right)) is similar to that of an alert with the only difference being that an event record captures both the reported date (i.e., the date of newspaper publication) and the event date (i.e., when the newspaper article reports the protest as having happened). The GSR is available from Nov 2012 and is used in this thesis primarily to help evaluate the performance of our system. A manual examination of GSR (as mentioned in Section 1) revealed that over 75% of the protests were organized and had clear triggering circumstances with political entrepreneurs leading the charge to protest.

3.1.1 Lead Time vs Accuracy of Forecast Date

It is important to understand which alerts *can* be matched to specific events. Note that there are four dates in an (alert,event) combination (see Fig. 3.2):

1. The date the forecast is made (*forecast date*)
2. The date the event is predicted to happen (*predicted event date*)
3. The date the event actually happens (*event date*)
4. The date the event is reported in a GSR source (*reported date*)

For an event to be qualified as having been predicted by a warning, *forecast date* < *reported date* (recall that time is measured in granularities of days). The *lead time* is given as (*reported date* − *forecast date*), i.e., the number of days by which we 'beat the news.' In constrast, the difference between *predicted event date* and *event date*, i.e., |*event date* −

predicted event date|. is one of *quality* or accuracy. Ideally we require *lead time* to be as high as possible and $|event\ date - predicted\ event\ date|$ to be as low as possible.

3.1.2 Other Quality Aspects

Forecasting the event date accurately is only one aspect of quality. Recall that alerts also forecast the location. A scoring formula is also defined for the location quality and overall quality score is defined as a sum over both the date and location scores.

$$\text{Quality score } (QS) = (DS + LS) * 2$$

where DS and LS denote the date score and location score, respectively. Each of these scores is in turn defined next:

$$DS = 1 - \min(|event\ date - predicted\ event\ date|, 7)/7$$

If the date of the event listed in the warning is the same as the actual date of the event, then DS is 1. On the other hand, if these dates are farther than 7 days apart, then DS is 0.

Location score (LS) can be defined in many ways. Because location is defined in terms of triples of (country, state, city), one approach is to use a tiered formula. Comparing a GSR event with a warning, we can obtain a score triple of (l_1, l_2, l_3) where l_1 is the country-level score, l_2 is the state-level score, and l_3 is the city-level score. Each of these scores have a value of 0 if they do not match and 1 is they match. Then the match between submitted warning location and the GSR location is given by:

$$LS = \frac{1}{3}l_1 + \frac{1}{3}l_1l_2 + \frac{1}{3}l_1l_2l_3$$

An alternative way to define location score is as:

$$LS = (1 - \min(\text{dist}, 300)/300)$$

where dist denotes the distance (in km) between the city predicted and the GSR city. All city location names are standardized to the World Gazeteer which provides latitude and longitude values, thus facilitating the computation of distance. We use the physical distance based criteria for our purposes.

3.1.3 Inclusion Criteria

Thus far we have demonstrated, given a warning-event(W-E) pair, how we can score their fitness. Inclusion criteria define which W-E pairs *can* even be considered for scoring. We have already mentioned one inclusion criterion, viz. that lead time must be > 0. The full list of inclusion criteria we will consider are:

1. Lead time > 0
2. Both warning and event are for the same country.
3. The *predicted event date* and *event date* must be within 7 days of each other.

A fourth, optional (and stringent), criterion we will use is:

4. Both predicted location and event location must be within 300km of each other.

It is important to distinguish the inclusion criteria from the scoring criteria. Inclusion criteria define which W-E pairs are allowable. Scoring criteria determine, from these allowable W-E pairs, what their score will be.

3.1.4 Matching Alerts to Events

Thus far we have assumed that we matching an alert to a GSR event. In practice, the problem is we are given a set of issued alerts and a set of GSR events and we must determine the quality of the match: which alert would correspond to which event? One strategy is to construct a bipartite graph between the set of alerts and the set of events, where allowable edges are those that satisfy the inclusion criteria, and where weights on these allowable edges denote their quality scores. We then construct a maximum weighted bipartite matching, e.g., see Fig. 3.3. Such matchings are conducted on a monthly basis with a lookback period to bring in unmatched warnings from the previous month.

The maximum weighted bipartite matching is done using Hungarian algorithm [35]

3.2 Probabilistic Soft Logic

In this section, we briefly describe probabilistic soft logic (PSL) [36], a key component of our geocoding strategy described later. PSL is a framework for collective probabilistic reasoning on relational domains. PSL models have been developed in various domains, including collective classification [37], ontology alignment [38], personalized medicine [39], opinion diffusion [40] , trust in social networks [41], and graph summarization [42]. 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

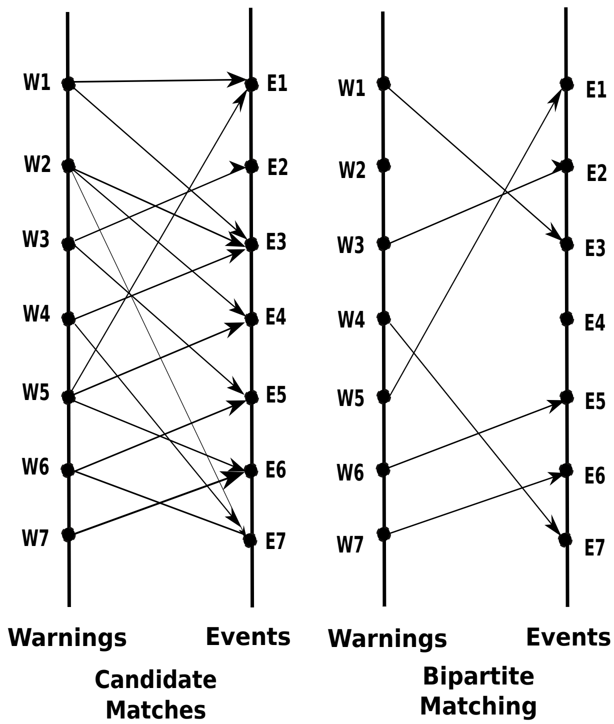


Figure 3.3: Given a set of candidate warning-event matches (left), we evaluate the performance of EMBERS using either a regular bipartite matching (right).

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 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 sat-*

isfaction. A rule $r \equiv r_{body} \rightarrow r_{head}$ is satisfied if and only if the truth value of head is at least 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

$$f(I) = \frac{1}{Z}\exp[-\sum_{r \in R} \lambda_r (d_r(I))^p] \tag{3.1}$$

$$Z = \int_I \exp[-\sum_{r \in R} \lambda_r (d_r(I))^p] \tag{3.2}$$

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.

Chapter 4

Approach

The general approach we adopt is to identify open-source documents that appear to indicate civil unrest event planning, extract relevant information from identified documents and use that as the basis for a structured warning about the planned event (see Fig. 4.1). We ingest a wide array of textual documents, including RSS feeds (news and blogs), mailing lists, URLs referenced in tweets, the contents of the tweets themselves, and Facebook event pages. All harvested documents are subjected to linguistic analysis; candidate documents are identified using a list of (learnt) phrases associated with protest event planning; date and location information is extracted from the text and reasoned about to generate a warning. Location information is standardized to conform to a standard (in our case, we use the World Gazeteer). Each of these processing steps (see Fig. 4.1) is outlined next.

4.1 Data Ingest

RSS Feeds (News/Blogs)

We ingest a total of 9498 feeds of which 6236 are news and the rest are blogs. The list of news sources to ingest was mainly obtained from the Latin American Network Information Center (LANIC) webpage. Some other online sources like Wikipedia, `www.onlinenewspapers.com`, `www.w3newspapers.com`, etc., were used to find more sources of online news in Latin America. Similarly in order to obtain a list of blogs to follow we used blog search engines like `www.technorati.com`.

We then subscribed to the RSS feeds of all identified sources. For sources, that did not have a feed of its own, we used Google Feed. We can also use RSS feed generators like `www.feedity.com` for the same purpose. We used Google Feed API to collect all the RSS feeds until its shutdown in july 2013. After that we moved to a custom feed collection solution in python that made use of the feedparser library (`http://code.google.com/p/`

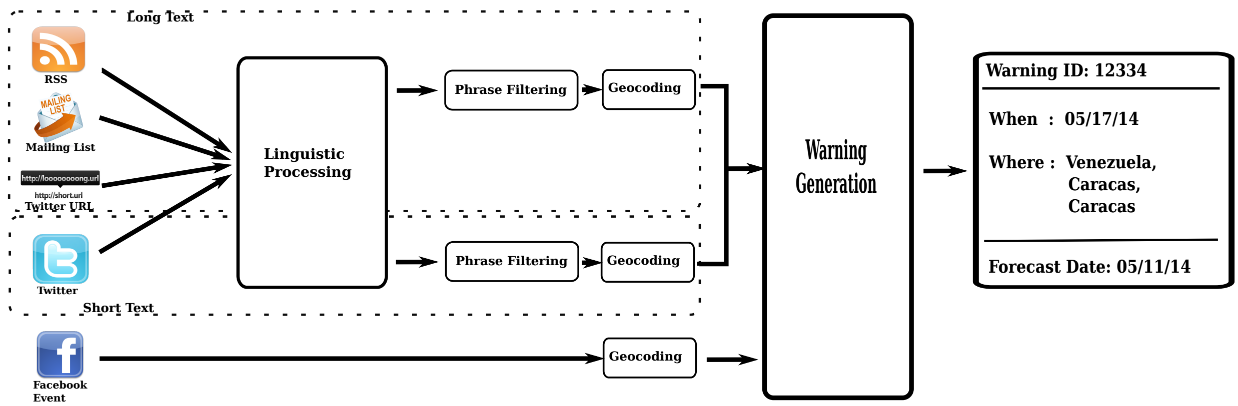


Figure 4.1: Schematic of the planned protest detector that ingests five different types of data sources.

`feedparser`).The Fig. 4.2 shows the number of messages that we ingest for each country on a daily basis.

Twitter

We ingest twitter making use of Datasift's Twitter collection engine. Datasift provides the ability to query and stream tweets in real time. These tweets are augmented with various types of metadata including the user profile of the tweeting user or geotagged attributes and the query can target any of these. Targeting tweets that come from a particular geographic area, e.g. Latin America, can be tricky. While some tweets use geotags to specify the location of the tweet, these tweets only comprise about 5% of the total number of tweets and may not be representative of the population overall (i.e. geotagged tweets come from people who have smart phones who also tend to be more affluent). Therefore, it is important to use other information to build a query that targets relevant tweets. In building our query we consider geotag bounding boxes (structured geographical coordinates), Twitter Places (structured data), user profile location (unstructured, unverified strings), and finally mentions of a location contained in the body of the tweet.

Facebook Events

We use Facebook Graph API and Facebook Query Language (FQL) to obtain Facebook Event data. The Graph API is used to query/search for Facebook Events with a particular keyword. We make use of a very specific list of 20 keywords that signify protest or ways to protest in one of three prominent languages – English, Spanish and Portuguese – in our region of interest. The Graph API returns a list of all Facebook Event IDS that contain a keyword irrespective of its location. The IDs of Event Pages are then used in a FQL

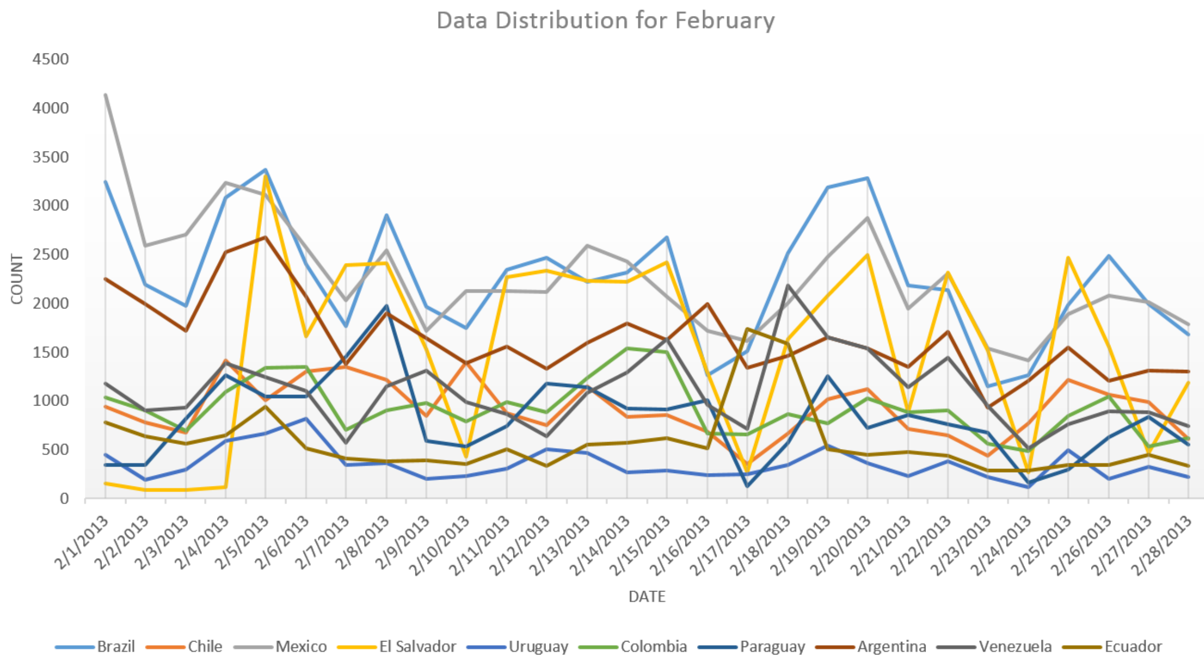


Figure 4.2: Distribution of RSS feeds across countries

query to obtain the features of an event like count of members, creator of the event, number of invitations sent, number of accepted invitations, number rejected, location and venue, start and end date, description of event, host details,etc. The host details in an Event page and the details about its creator are very critical in identifying if the Event page is protest related. The Facebook-Event data we collected spans from december 2013 to march 2014. Obtaining historical data from Facebook depends on whether the Facebook Event page created sometime in the history is still kept active by the creator. If the page is deleted then it is not possible to obtain the data.

Besides filtering the Facebook Event dataset for the presence of certain keywords we also do some filtering based on origin/creator of an event. We ignore events hosted by musical bands, restaurants or other commercial organizations. We also ignore events hosted by an individual. The list of organizations to ignore was built manually from the list of possible values as observed. Fig. 4.3 shows an example of a Facebook Event page calling for protest against the construction of the Temple of Solomon in Sao Paulo, Brazil.

4.2 Linguistic Preprocessing

All textual input (e.g., tweets, news articles, blog postings) is subjected to shallow linguistic processing prior to analysis. This involves identifying the language of the document, distinguishing the words (tokenization), normalizing words for inflection (lemmatization), and

Figure 4.3: Example of a Facebook Event

identifying expressions referring to people, places, dates and other entities and classifying them (named entity extraction). Since our region of interest is Latin America, the collection of text harvested is inherently multilingual, with Spanish, Portugese, and English as the dominating languages; we use Basis Technology's Rosette Linguistics Platform (RLP) suite of multilingual commercial tools (`http://www.basistech.com/text-analytics/rosette/`) for this stage. The output of linguistic preprocessing serves as input to subsequent deeper analysis in which date expressions are normalized and the geographic focus of the text identified.

4.3 Temporal Information Extraction

Temporal Information Extraction(TIE) is particularly crucial to the identification of future oriented statements. We use the TIMEN [18] date normalization package to normalize and deindex temporal expressions. A temporal expression is any word or phrase that describe a point, duration or recurrence in time. TIMEN is a community-driven tool for temporal expression normalisation .i.e, interpretting a temporal expression accurately and converting it into ISO-9601 datetime stamp. TIMEN allows a good amount of reusability and extensibility. TIMEN architecture also clearly separates the algorithms which do the normalization from the knowledge and rule base necessary for the process, thus making it wasy to extend to multiple languages. The TIMEN system makes use of meta-data such as the day of publication, and other information about the linguistic context of the date expression to determine for each date expression, what day (or week, month or year) it refers to. For example in a tweet produced on June 10, 2014, the occurrence of the term *Friday* used in a future-tense sentence *We'll get together on Friday* will be interpreted as June 13, 2014. Each

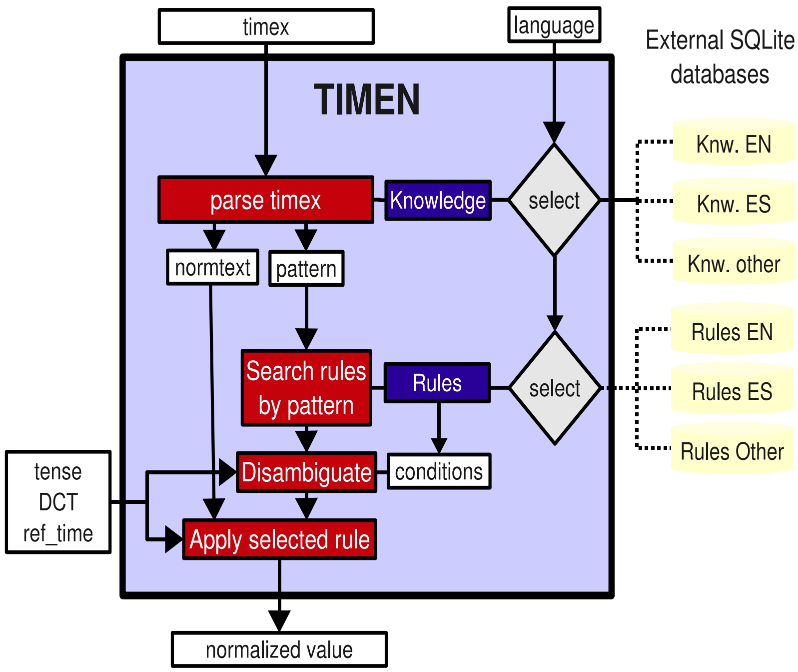


Figure 4.4: TIMEN architecture

expression identified as a date by the RLP preprocessor is normalized in this way.Internally, TIMEN converts a temporal expression to be normalized into a symbolic representation using a knowledge base (KB). The rules are then matched against this representaion to produce a normalised output.The TIMEN architecture is shown in 4.4

4.4 Phrase filtering

In order to identify relevant documents, input documents are filtered on a set of key phrases, i.e., the text of the document is searched for the presence of one or more key phrases in a list of phrases which are indicative of an article's focus being a planned civil unrest event. The list of key phrases indicating civil unrest planning was obtained in a semi-automatic manner, as detailed in Section 4.5. Articles which do match are processed further, those that do not are ignored.

4.4.1 Phrase matching

Our key phrase matching is highly general and linguistically sophisticated. The phrases in our list are general rules for matching, rather than literal string sequences. Typically a phrase specification comprises: two or more word lemmas, a language specification, and a separation threshold. This indicates that words—potentially inflected forms—in a given sequence potentially separated by one or more other words, should be taken to be a match. We determined that this kind of multi-word key phrases was more accurate than simple keywords for extracting events of interest from the data stream.

The presence of a keyphrase is checked by searching for the presence of individual lemmas of the keyphrase within the same sentence separated by at most a number of words that is fewer than the separation threshold. This method allows for linguistically sophisticated and flexible matching, so, for example, the keyphrase [***plan protest*, 4, English**] would match the sentence *The students are planning a couple big protests tomorrow* in an input document.

4.5 Phrase list development

The set of key phrases was tailored (slightly) to the genre of the input. In particular different phrases were used to identify relevant news articles and blogs from those used to filter Tweets. The lists themselves were generated semi-automatically.

Initially, a few seed phrases were obtained manually with the help of subject matter experts. An analysis of news reports for planned protests in 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 combined these with future-oriented verbs, e.g., *to organize*, *to prepare*, *to plan*, and *to announce*. For twitter, shorter phrases were identified, and these had a more direct call for action, e.g., *marchar*, *manhã de mobilização*, *vamos protestar*, *huelga*.

To generalize this set of phrases, the phrases were then parsed using a dependency parser [43] and the grammatical relationship between the core nominal focus word (e.g., *protest*, *manifestación*, *huelga*) and any accompanying word (e.g., *plan*, *call*, *anunciar*) was extracted. These grammatical relations were used as extraction patterns as in [13] to learn more phrases from a corpora of sentences extracted from the data stream of interest (either news/blogs or tweets). This corpus consists of sentences that contained any one of the nominal focus words and also had mentions of a future date. The separation threshold for a phrase was also learned, being set to the average number of words separating the nominal focus and the accompanying word.

The set of learned phrases is then reviewed by a subject matter expert for quality contraol. Using this approach, we learned 112 phrases for news articles and blogs and 156 for tweets. This phrase learning process is illustrated in Fig. 4.5.

4.6 Geocoding

After linguistic preprocessing and suitable phrase filtering, messages are geocoded with a specification of the geographical focus of the text—specified as a city, state, country triple—that indicates the locality that the text is about. We make use of different geocoding methodologies for Twitter messages, for Facebook Events pages, and for news articles and blogs.

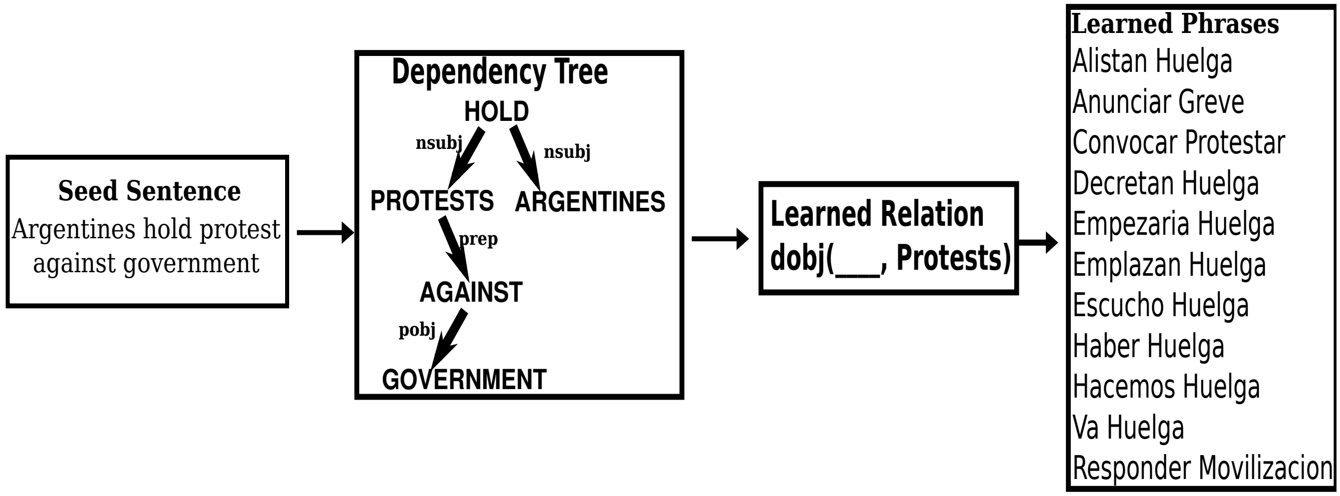


Figure 4.5: An example of phrase learning for detecting planned protests.

These are described below.

4.6.1 Twitter

A lot of work has been done regarding geocoding of twitter text content.Different types of location information like location from where the tweet has been written, content location, user's home location etc. For tweets, the geo-focus of the message is generated by a fairly simple set of heuristics. In particular, Twitter geocoding is achieved by first considering the most reliable but least available source, the geotag (latitude, longitude) of the tweet itself (this is available for about 10% of our sample from Twitter). This provide an exact geographic location that can be reverse geocoded into a place name and used as the geo-focus. We find the nearest geo-coded point in our extended gazetteer (using a kd-tree data structure) for this purpose. If the tweet is not geocoded, we consider Twitter *places* metadata and use place names present in these metadata fields to geocode the place names into geographical coordinates. Finally, if none of this is available, 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. Additional toponym disambiguiation heuristics are used to identify the actual referent of the mention.

4.6.2 Facebook

Similar methods are used to geocode event data extracted from Facebook event pages. Since only Facebook events that have a venue are used and since the venue of a Facebook event generally specifies a latitude, longitude, and physical address information, identifying the location is a fairly trivial task. In cases where only latitude and longitude are given, we apply reverse-geocoding mechanisms similar to those used for Twitter.

4.6.3 News and Blogs

For longer articles such as news articles, the geo-focus of the message is identified using much more complex methods To extract the protest location from news articles, we use PSL to build probabilistic models that infer the intended location of a protest by weighing evidence coming from the Basis entity extractions and information in the World Gazeteer.

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 article, 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.

$$
\begin{aligned}
&ENTITY(L, location) \tilde{\wedge} REFERSTO(L, locID) \\
&\qquad \rightarrow PSLLOCATION(Article, locID)
\end{aligned}
$$

$$
\begin{aligned}
&ENTITY(C, location) \tilde{\wedge} IsCountry(C) \\
&\qquad \rightarrow ArticleCountry(Article, C)
\end{aligned}
$$

$$
\begin{aligned}
&ENTITY(S, location) \tilde{\wedge} IsState(S) \\
&\qquad \rightarrow ArticleState(Article, S)
\end{aligned}
$$

(Note that the above are not deterministic rules; e.g., they do not use the logical conjunction $\wedge$ but rather the Lukasiewicz t-norm based relaxation $\tilde{\wedge}$. Further, these rules do not fire deterministically but are instead simultaneously solved for satisfying assignments as described in Section 3.2.)

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.

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

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

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

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. As an example of how PSL aids in location identification, the example from Fig. 1.1 is revisited in Fig. 4.6.

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

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

4.7 Warning Generation

After being subject to the preprocessing steps, above, all documents that are identified as containing a key phrase are further filtered by searching for the presence a future date in the passage containing the key phrase and for the existence of an identified geographical focus for the text. Documents that meet all these critera are used as the basis for a warning about a planned civil unrest event (Twitter postings are only used as the basis for a warning if the tweet is re-tweeted at least five times). A warning is generated for the date indicated by the future date expression and the location which is the geographical focus of the document. In the case of Facebook, an event page is considered to be a good evidence for an alert if there are more attendees (sum of number of accepted invitations and number of 'maybe') for the event than rejects (number of declined invitations). The date and location are read off from the event page directly.

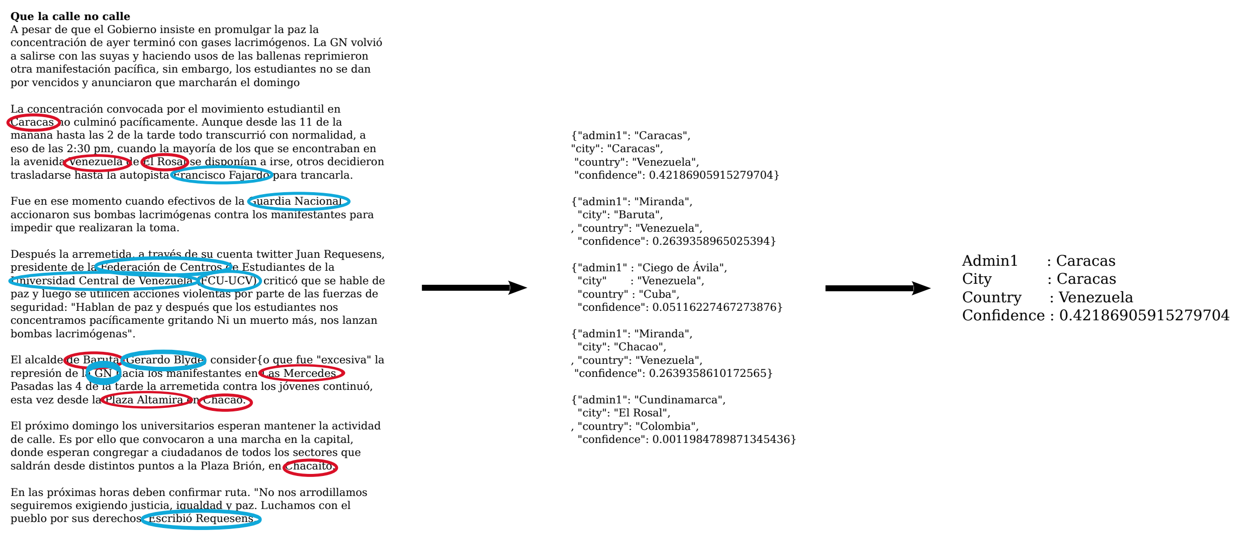


Figure 4.6: Red circles denote named entities identified as locations and blue denotes other types of entities. The article describes students planning a march on Sunday. It identifies multiple locations, e.g., Chacao, El Roso, and the Francisco Fajardo highway where protests have been happening. There is also a reference to a quote by the mayor of Baruto. Mentions of such multiple locations are resolved using our PSL program to the intended location, here Caracas.

Chapter 5

Experiments

We evaluate our planned protest detection system using metrics similar to those described by Ramakrishnan et al. [23] in evaluating their work. Given a set of alerts issued by the system and the GSR comprising actual protest incidents, we aim to identify a correspondence between the two sets via a bipartite matching. An alert can be matched to a GSR event only if i) they are both issued for the same country, ii) the alert's predicted location and the event's reported location are within 300km of each other (the distance offset), and iii) the forecasted event date is within a given interval of the true event date (the date offset). Once these inclusion criteria apply, the quality score (QS) of the match is defined as a combination of the location score (LS) and date score (DS):

$$\text{QS} = (LS + DS) * 2 \tag{5.1}$$

where

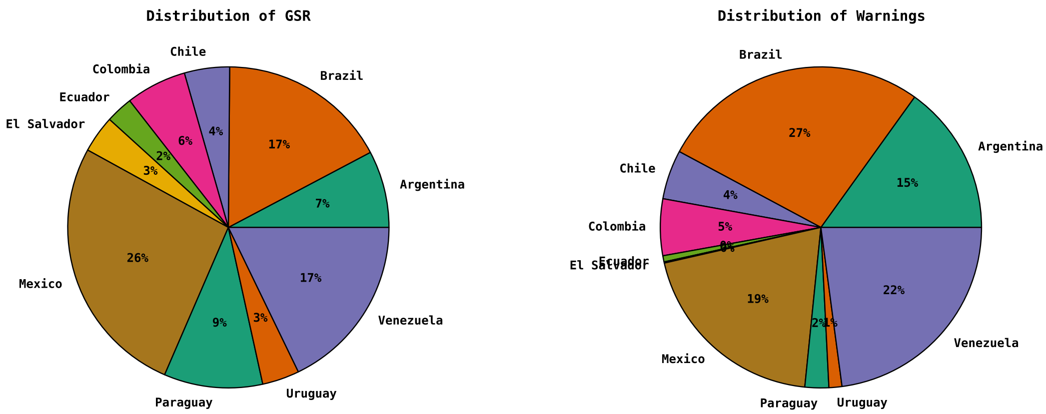
$$\text{LS} = 1 - \frac{\min(\text{distance offset}, 300)}{300} \tag{5.2}$$

and

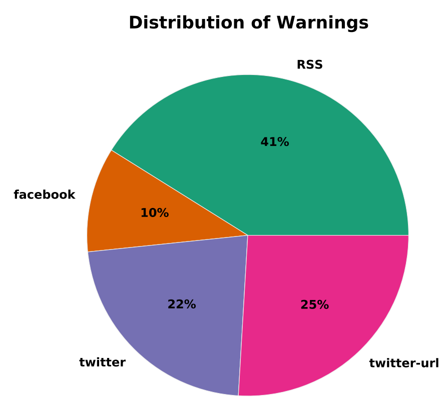
$$\text{DS} = 1 - \frac{\min(\text{date offset}, \text{INTERVAL})}{\text{INTERVAL}} \tag{5.3}$$

Here, we explore INTERVAL values from 0 to 7. if an alert (conversely, GSR event) cannot be matched to any GSR event (alert, respectively), these unmatched alerts (and events) will negatively impact the precision (and recall) of the system. The lead time, for a matched alert-event pair, is calculated as the difference between the date on which the forecast was made and the date on which the event was reported (this should not be confused with the date score, which is the difference between the predicted event date and the actual event date). Lead time concerns itself with reporting and forecasting, whereas the date score is concerned with quality or accuracy.

We conduct a series of experiments to evaluate the performance of our system.



(a) Distribution of GSR events (left) and Alerts(right) by country



(b) Distribution of alerts from different sources.

Figure 5.1: Distribution of alerts and GSR events

Table 5.1: QS=Quality Score; Pr=Precision; Rec=Recall; LT=Lead Time. AR=Argentina; BR=Brazil; CL=Chile; CO=Colombia; EC=Ecuador;SV=El Salvador; MX=Mexico; PY=Paraguay; UY=Uruguay; VE=Venezuela. A − indicates that the source did not produce any warnings for that country in the studied period.

	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

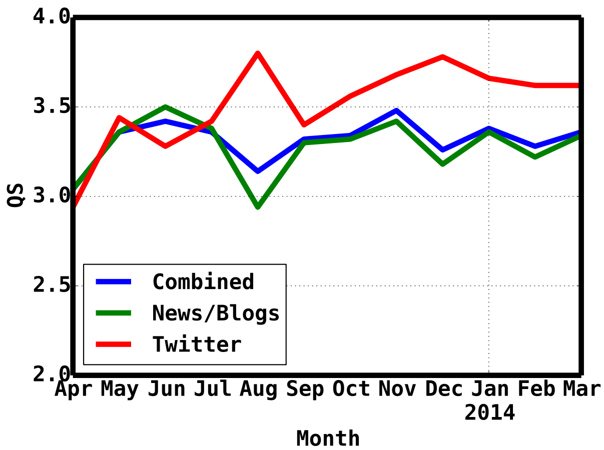


Figure 5.2: Quality Score over the months

Distribution of protests detected by the system compared with the actual distribution of protests in the GSR? Fig. 5.1a reveals pie charts of both distributions. As shown, Mexico, Brazil, and Venezuela experience the lion's share of protests in our region of interest, and the protests detected also match these modes although not the specific percentages. The smaller countries like Ecuador, El Salvador, and Uruguay do experience protests but which are not as prominently detected as those for other countries; we attribute this to their smaller social media footprint (relative to countries like Brazil and Venezuela). 5.1 reveals the distribution of warnings from different sources.

Are there country-specific selective superiorities for the different data sources considered here? Table 5.1 presents a breakdown of perfomance, country-wise and source-wise, of our approach for a recent month, viz. March 2014. It is clear that the multiple data sources are necessary to achieve a high recall and that by and large these sources are providing mutually exclusive alerts. (Note also that some data sources do not produce alerts for specific countries.) Between Twitter and Facebook, the former is a better source of alerts

Table 5.2: AR=Argentina; BR=Brazil; CL=Chile; CO=Colombia; EC=Ecuador;SV=El Salvador; MX=Mexico; PY=Paraguay; UY=Uruguay; VE=Venezuela. A − indicates that the source did not produce any warnings for that country in the studied period.

Source		AR	BR	CL	CO	EC	SV	MX	PY	UY	VE	All
News/Blogs	LS	0.82	0.76	0.75	0.60	-	**0.75**	0.66	0.79	**0.79**	**0.95**	0.81
	DS	0.75	0.81	0.78	0.77	-	**0.86**	0.82	0.90	**0.83**	**0.95**	0.86
Facebook	LS	**1.0**	**0.92**	-	**1.00**	-	-	**0.86**	**0.98**	-	-	**0.93**
	DS	0.85	**0.89**	-	**1.00**	-	-	**1.00**	**1.00**	-	-	0.90
Twitter	LS	0.88	-	**0.84**	0.81	**0.45**	-	0.71	**0.98**	-	0.91	0.89
	DS	**0.88**	-	**0.92**	0.84	**0.71**	-	0.86	0.94	-	0.93	**0.92**

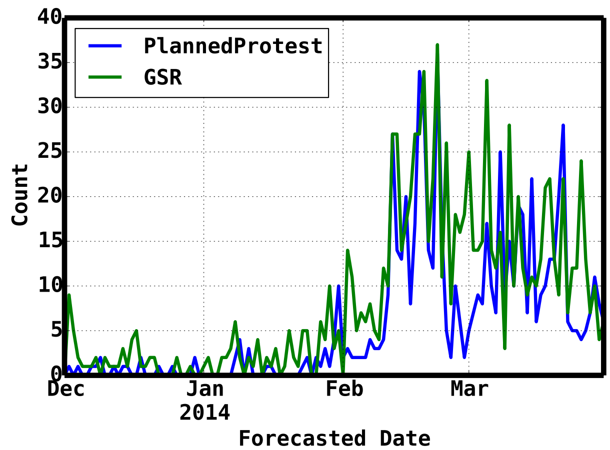


Figure 5.3: Venezuelan Protests

for countries like Chile and the latter is a better source for Argentina, Brazil, Colombia, and Mexico. News and blogs achieve higher recall than social media sources indicating that most plans for protests are announced in established media. They are also higher quality sources for alerts in countries like El Salvador, Paraguay, and Uruguay. Finally, note that news and blogs offer a much higher lead time (4.57 days) as compared to that for Facebook (2.44 days) or for Twitter (2.82 days). The quality scores are further broken down in Table 5.2 into their date and location components. A longitudunal perspective on quality scores is given in Fig. 5.2. Note that in general Twitter tends to have a higher quality score as multiple re-tweets of future event mentions is a direct indicator of the popularity of an event as well as the intent of people to join an event. In contrast, mentions of future events in news do not directly shed any insight into popularity or people's support for the event's causes.

How did our system fare in detecting key country-wide 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 evalua-

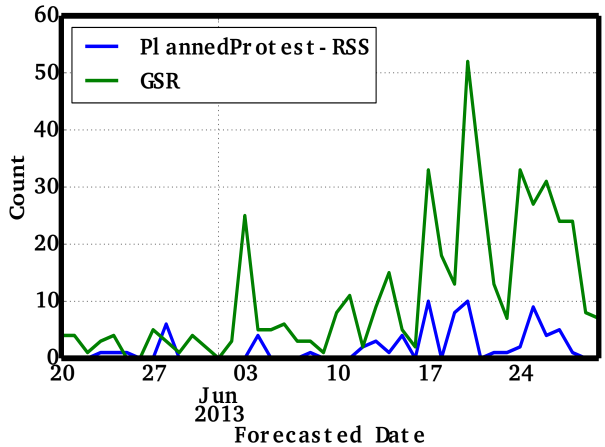


Figure 5.4: System Performance during Brazilian Spring

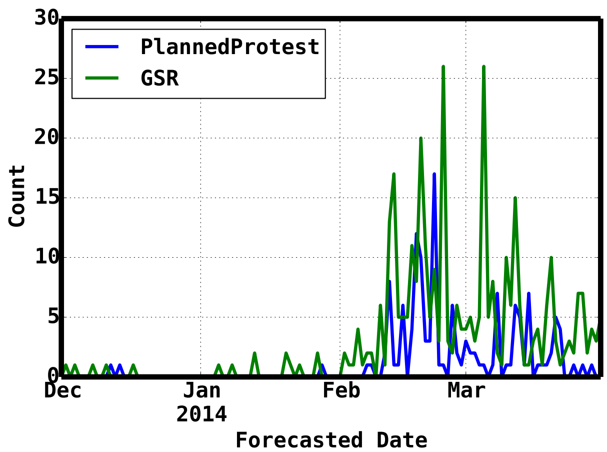


Figure 5.5: System Performance during Brazilian Spring

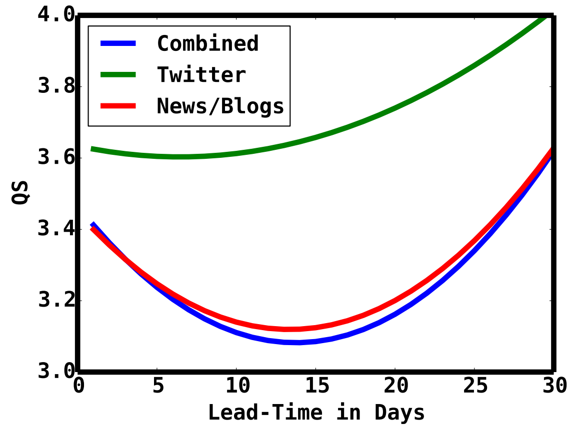


Figure 5.6: Lead-Time vs Quality Score

tion. Fig. 5.3 and Fig. 5.4 describe our performance under these two situations illustrating the count of protests detected against the GSR. Notice that our system was able to identify the Venezuelan protests much better than the Brazilian protests. This is because there was a significant amount of spontaineity to the Brazilian protests; they arose as discontent about bus fare increases but later morphed into a broader set of protests against government and most of these subsequent protests were not planned.

Fig. 5.5 shows how our system was successful in predicting the violent protests also in venezuela.An alert is considered to be for a violent or not using a binary text based naive bayes classifier.

What is the tradeoff between lead time and quality? Fig. 5.6 shows that the QS of the planned protest model decreases (as expected) with lead time, initially, but later rises again. The higher quality scores toward the right of Fig. 5.6 are primarily due to Facebook event pages.

How does the method perform under stringent matching criteria? Fig. 5.7 shows the perfomance of the model when the matching window is varied from 7 to 1 in steps. We can see that the performance degrades quite gracefully even under the strict matching interval of a 1-day difference.

What is the distribution of quality scores? The clear mode toward the right side of the Fig. 5.8 signifies that a majority of the planned protest alerts are of high quality. Further, the quality score distribution is unimodal suggesting that the careful reasoning of locations and date normalization are crucial to achieving high quality.

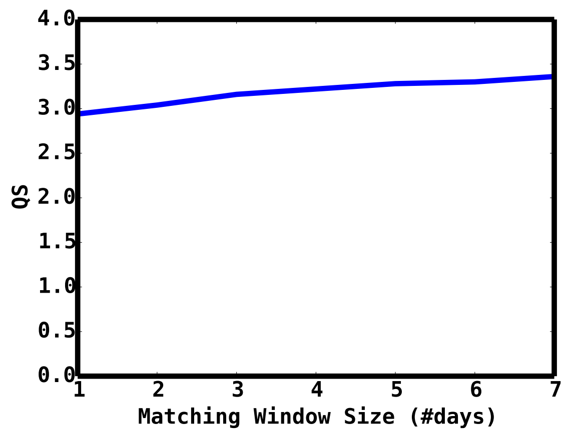


Figure 5.7: QS vs Matching Interval Trade-Off

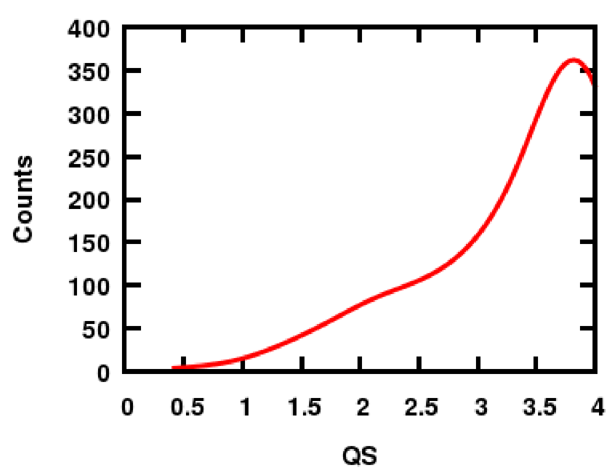


Figure 5.8: Quality Score Distribution

Chapter 6

Discussion

We have described an approach to forecasting protests by detecting mentions of future events in news and social media. The two twin issues of i) resolving the date and ii) resolving the location have been addressed satisfactorily to realize an effective protest forecasting system. As different forms of communication media gain usage, systems like ours will be crucial to understanding the concerns of citizenry.

Our future work is aimed at three aspects. First, to address situations such as nationwide protests and systems of protests, we must generalize our system from generating protests at a single article level to digesting groups of articles. This will require more sophisticated reasoning using PSL programs. Second, we would like to generalize our approach that currently does detection of overt plans for protest to not-so-explicitly stated expressions of discontent. Finally, we plan to consider other population-level events of interest than just civil unrest, e.g., domestic political crises, and design detectors to recognize the imminence of such events.

Bibliography

[1] 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*, ser. ICUIMC, 2010.

[2] 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*, ser. CIKM, 2011.

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

[4] 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*, ser. BNAIC, 2013.

[5] A. H. Hurriyetoglu, 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*, ser. DIR, 2013.

[6] 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.

[7] 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*, ser. ISI, 2013.

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

[9] 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.

[10] 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*, ser. WWW, 2004.

[11] 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*, ser. IJCAI, 2007.

[12] 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*, ser. HLT, 2011.

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

[14] 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*, ser. KDD, 2012.

[15] 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*, ser. WWW, 2010.

[16] M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, J. Moszkowicz, and J. Pustejovsky, "The tempeval challenge: identifying temporal relations in text," *Language Resources and Evaluation*.

[17] J. Pustejovsky, J. M. Castano, R. Ingria, R. Sauri, R. J. Gaizauskas, A. Setzer, G. Katz, and D. R. Radev, "Timeml: Robust specification of event and temporal expressions in text." *New directions in question answering*.

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

[19] I. Mani and G. Wilson, "Robust temporal processing of news," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ser. ACL, 2000.

[20] 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*, ser. WSDM, 2013.

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

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

[23] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz *et al.*, "'beating the news' with embers: Forecasting civil unrest using open source indicators," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.

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

[25] 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*, ser. WSDM, 2012.

[26] 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*, ser. ICWSM, 2013.

[27] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham, "Inferring private information using social network data," in *Proceedings of the 18th International Conference on World Wide Web*, ser. WWW, 2009.

[28] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: improving geographical prediction with social and spatial proximity," in *Proceedings of the 19th international conference on World wide web*, ser. WWW, 2010.

[29] E. Amitay, N. Har'El, R. Sivan, and A. Soffer, "Web-a-where: geotagging web content," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2004, pp. 273–280.

[30] C. Fink, C. D. Piatko, J. Mayfield, T. Finin, and J. Martineau, "Geolocating blogs from their textual content." in *AAAI Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0*, 2009.

[31] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, ser. CIKM, 2010.

[32] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, "Geographical topic discovery and comparison," in *Proceedings of the 20th international conference on World wide web*, ser. WWW, 2011.

[33] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, "Towards social user profiling: unified and discriminative influence model for inferring home locations," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD.

[34] R. Li, S. Wang, and K. C.-C. Chang, "Multiple location profiling for users and relationships from social network and content," *Proceedings of the VLDB Endowment*, 2012.

[35] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, 1955.

[36] 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.

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

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

[39] 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.

[40] 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.

[41] 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.

[42] A. "Memory, A. Kimmig, S. H. Bach, L. Raschid, and L. Getoor, """graph summarization in annotated data using probabilistic soft logic"," in *"Proceedings of the International Workshop on Uncertainty Reasoning for the Semantic Web"*, ser. URSW, "2012".

[43] L. Padró and E. Stanilovsky, "Freeling 3.0: Towards wider multilinguality," in *Proceedings of the Language Resources and Evaluation Conference*, ser. LREC, 2012.