User Interfaces for an Open Source Indicators Forecasting System

Nathan W. Self

Thesis 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
Chris North
Kurt Luther

August 13, 2015
Blacksburg, Virginia

Keywords: Visualization, Forecasting, Intelligence Analysis, Open Source Indicators.

User Interfaces for an Open Source Indicators Forecasting System

Nathan W. Self

(ABSTRACT)

Intelligence analysts today are faced with many challenges, chief among them being the need to fuse disparate streams of data and rapidly arrive at analytical decisions and quantitative predictions for use by policy makers. A forecasting tool to anticipate key events of interest is an invaluable aid in helping analysts cut through the chatter. We present the design of user interfaces for the EMBERS system, an anticipatory intelligence system that ingests myriad open source data streams (e.g., news, blogs, tweets, economic and financial indicators, search trends) to generate forecasts of significant societal-level events such as disease outbreaks, protests, and elections. A key research issue in EMBERS is not just to generate high-quality forecasts but provide interfaces for analysts so they can understand the rationale behind these forecasts and pose why, what-if, and other exploratory questions.

This thesis presents the design and implementation of three visualization interfaces for EMBERS. First, we illustrate how the rationale behind forecasts can be presented to users through the use of an 'audit trail' and its associated visualization. The audit trail enables an analyst to drill-down from a final forecast down to the raw (and processed) data sources that contributed to the forecast. Second, we present a forensics tool called Reverse OSI that enables analysts to investigate if there was additional information either in existing or new data sources that can be used to improve forecasting. Unlike the audit trail which captures the transduction of data from raw feeds into alerts, Reverse OSI enables us to posit connections from (missed) forecasts back to raw feeds. Finally, we present an interactive machine learning approach for analysts to steer the construction of machine learning mod-

els. This provides fine-grained control into tuning tradeoffs underlying EMBERS. Together, these three interfaces support a range of functionality in EMBERS, from visualization of algorithm output to a complete framework for user feedback via a tight human-algorithm loop. They are currently being utilized by a range of user groups in EMBERS: analysts, social scientists, and machine learning developers, respectively.

Supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC00337, the U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBA, or the U.S. Government.

Acknowledgments

I would like to thank my advisor, Naren Ramakrishnan, for helping me make sense of this crazy world of research. To my committee members, Chris North and Kurt Luther, I am indebted for their help with shaping this thesis into a more compelling argument. I would also like to thank the entire EMBERS team, and in particular coding guru Patrick Butler, for their work on forecasting and for late nights preparing for site visits. For help on the Reverse OSI interface and helping analyze the results of the study, I would like to thank Kristen Summers, David Mares, and Parang Saraf. Finally, to my family and friends who were full of support and patience I express deep gratitude.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Intelligence analysts today are faced with many challenges, chief among them being the need to fuse disparate streams of data and rapidly arrive at analytical decisions and quantitative predictions for use by policy makers. A forecasting tool to anticipate key events of interest is an invaluable aid in helping analysts cut through the chatter.

1.1 EMBERS

Our team is a universityindustry partnership developing advanced forecasting algorithms for significant societal events such as disease outbreaks, elections, domestic political crises, and civil unrest incidents. The system generated by our effort (EMBERS, for Early Model Based Event Recognition using Surrogates) is an automated environment to ingest myriad data streams and process them into alerts (or forecasts) about population-level events of interest. The scope of EMBERS spans several countries of Latin Americanamely, Argentina, Bolivia, Brazil, Chile, Costa Rica, Colombia, Ecuador, El Salvador, French Guiana, Guatemala, Honduras, Mexico, Nicaragua, Paraguay, Panama, Peru, Uruguay, and Venezuela. Our

team includes researchers in data mining, machine learning, natural language processing, network dynamics, computational epidemiology, political science, systems integration, and Latin American studies.

The EMBERS-generated forecasts are fine-grained in that they qualify the who, why, where, and when of an event. For instance, "Teachers will protest for wage-related reasons in the city of Curitaba, Brazil, this coming Wednesday" is an example of an alert. Forecasting the dates, locations, and participating populations in this manner can offer situational awareness into unfolding events. In addition, aggregating this information and the data that supports it can offer insights into the broader sociocultural environment. For example, an analyst who sees an increase in protests in a given population might examine the source data and find that certain ongoing issues, such as crime rates, are starting to produce more specific unrest than in the past, which in turn would spur analysis and insights of the factors affecting the events.

EMBERS has been generating alerts continuously since November 2012 without a human in the loop, as is the requirement of the Intelligence Advanced Research Projects Activity (IARPA) Open Source Indicators (OSI) program supporting the development of EMBERS. Unlike retrospective studies of predictability, alerts generated by EMBERS are emailed in real time to IARPA and must precede the event being forecast to count as a prediction. The received alerts are evaluated monthly by an independent test and evaluation team (MITRE). Analysts and subject matter experts at MITRE survey international and domestic newspapers of record in each country that EMBERS studies and catalog a master set of events in these countries, known as the gold standard report (GSR). This GSR is then matched against the forecasts generated by EMBERS leading to several evaluation measures: precision, recall, lead time, average quality score, and average probability score (confidence).

1.2 EMBERS Interfaces

A variety of data sources are integrated in EMBERS including mainstream and social media, and other indirect indicators of societal stability. A key feature in EMBERS is the harnessing of social media datasets (e.g., Twitter) which have been used as a weak predictor or as a correlative surrogate for many real-world events such as box office earnings [1], flu case counts [2], and even stock prices [3]. This line of research is full of thorny problems in data analysis such as how to detect puppet accounts that represent special interests rather than public opinion, how to account for spread of information outside of the observable social media space, and how to geocode entries to determine whether users are first or second hand sources. The above problems are exacerbated as we integrate social media data with other sources of data.

In developing EMBERS and deploying it in production, we have had to address the needs of several user groups: analysts who would like to understand the rationale behind a prediction, social scientists who would like to diagnose the cause for a missed forecast (false negative), and machine learning developers who would like to understand how to improve forecasting algorithms (and how). The latter problem is particularly important. Most algorithms for predictive analytics require fine tuning of parameters which can have significant impacts on the reliability and performance of a model. Even with a rough understanding of what such parameters mean, often the outcome of changing them is hard to predict. In some cases, parameter setting is best left to automated means, which further removes users from grasping what a model is doing and why. The group with the best skills for guiding a model, viz. experts in data science, does not necessarily intersect with the group that has the best skills for evaluating such a model's results, viz. experts and end users. This presents not only challenges for visualizing the processes underlying such systems but also for providing interfaces to interact with them.

1.3 Research Questions

One difficulty with forecasting systems such as EMBERS is that their outputs might succeed in terms of objective measures such as precision and recall but fail in terms of subjective measures such as interestingness, trustworthiness, and other measures of value or utility. This thesis presents the design and implementation of three interfaces that deal with different user interface aspects of EMBERS:

1. *EMBERS Audit Trail Visualizer:* Can we design an interface to explore the outputs of predictive models that operate on large quantities of data? Can it support exploration of why a forecast was generated? Can we support more complex questions like what if a data source is excluded?

2. *Reverse OSI:* Can we design an interface to investigate failures of machine learning algorithms? Can it scale to support many investigations of many failures?

3. *Interactive Model Building:* How can we leverage the domain expertise of users to improve the performance of EMBERS? Can we provide an interface that allows domain experts to build models that make predictions?

Chapter 2

EMBERS Audit Trail Visualizer

2.1 Requirements Analysis

To investigate our first research questions we built several interfaces for predictive models that are a part of the EMBERS system.

2.1.1 Background

As introduced earlier, EMBERS is a system for forecasting societal level events using open source indicators [4]. The system entails round-the-clock ingestion of raw data from multiple open source platforms (news and social media). These inputs undergo several stages of processing (enrichment) to compute additional attributes from the raw text data such as whether referenced events are in the past, present, or future and what geolocation is intended. These enriched messages are passed on to a suite of predictive models that make decisions about whether these data might be precursors to interesting events. When the EMBERS system decides that the level of precursors is high enough, it emits a forecast detailing several

attributes of the predicted event including time, place, and participants. This architecture is represented in Figure 2.1. Over the course of the EMBERS project, predicted events have included civil unrest in the form of protests and strikes, outbreaks of rare diseases, significant changes in stock market values, and weekly counts of reports of influenza like illnesses (ILI).

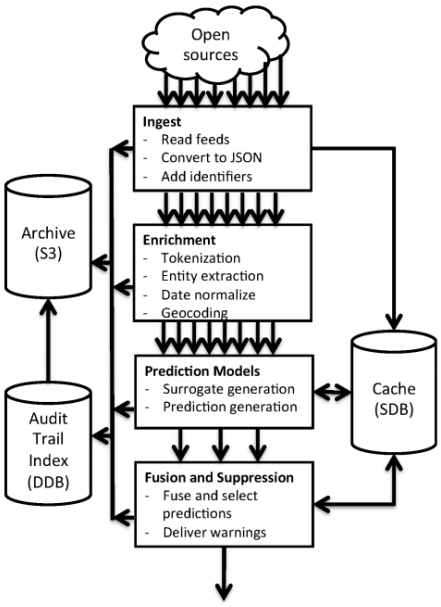


Figure 2.1: EMBERS system architecture.

Also, as introduced earlier, all forecasts made by the EMBERS system are submitted to a third party (MITRE) for evaluation [5]. As such, predictions must conform to a schema that defines what fields must be present in a forecast. For instance, forecasts for civil unrest events such as protests, riots, and strikes must include the location, the date on which the event will take place, the subgroup of the population that will take part, a code for the reason/rationale for the protest, and whether it will turn violent. As defined by the IARPA OSI program, an event is considered to be any gathering of people that is reported about in a news outlet after it has happened.

The models in EMBERS combine and synthesize large amounts of data to make forecasts.

For instance, predictions of influenza like illness (ILI) counts derive features from many data sources including historical weather data such as humidity and rainfall, historical counts of illness reports, the amount of recent Google searches for related words, and even Googles own prediction of influenza likelihood. Predictions generated by EMBERS are specific about what they think will happen but without supporting evidence from raw data it can be hard for analysts to interpret and evaluate them. To this end, we designed and implemented an interface for visualizing the key data transduction steps involved in the generation of EMBERS predictions. We refer to these steps as constituting an *audit trail*.

2.1.2 System Architecture

Each discrete unit (at any stage of EMBERS processing) is modeled as a message and is passed around by queues that connect different components of the system [6]. In the EMBERS system, input streams go through multiple enrichment and feature extraction steps before being passed to models for forecasting. For instance, as a tweet is ingested into the system it is packaged into a message with relevant details such as username, time of generation, and content. This message is passed through several main enrichment modules that add derived information. A natural language processing package from Basis Technology adds part of speech tags, language identification, and noun phrase detection. Messages that reference dates are parsed with the help of the TIMEN package [7] to convert relative date references (e.g., *tomorrow*, *next Friday*) with absolute dates. Sentiment analysis is done on each text using the ANEW [8] lexicon to generate a sentiment score measuring the valence, dominance, and arousal of tokens in that text.

A geocoder developed by the EMBERS team tackles the tricky task of determining what location a block of text is referring to. For tweets, a classifier determines the location from

the tweet content and also employs user profile or other geographic metadata available in the tweet payload. For longer texts, such as news articles, which might mention several locations, a more complex system using probabilistic soft logic determines location [9]. After enrichment, the tweet is given a unique index and stored so that it will be available for use by any model that is interested in using it. For instance, the Dynamic Query Expansion (DQE) model [10] will focus on all tweets from a country on a given day to forecast civil unrest. If DQE decides that an event will take place, it will send yet another message detailing the forecasted event.

At each stage of this process an audit trail is maintained to explain the provenance of each warning. Every time a new message is generated a "derivedFrom" field is added which includes a list of the unique IDs of every message that was an input for this messages generation. So in our case, our enriched tweet will have a "derivedFrom" field pointing to the raw tweet and the DQE warning message will have a "derivedFrom" field containing a list of all the IDs of enriched tweets it used. By following the trail of "derivedFrom" fields back to raw input, a complete collection of all the messages that were involved in the generation of a prediction can be organized. This audit trail is represented in EMBERS as a large, tree-like, hierarchical JSON object that can be traversed via "derivedFrom" tags.

2.1.3 Visualization Goals

Because audit trails can grow to include thousands of messages taking up hundreds of megabytes, it is difficult and impractical for anyone to analyze the provenance of a prediction by reading the audit trail alone. We developed both a dashboard to display the state of outstanding EMBERS forecasts and a series of visualizations that summarize the most important features of the audit trails for each predictive model. The goal of the dashboard

is to allow quick understanding of all the dimensions of a warning (event date, location, population, event type, etc.) and enable the user to assimilate and aggregate these dimensions. From the dashboard users can access detailed views of each forecast and visualizations of relevant parts of those forecasts' audit trails.

2.2 Design and Implementation

2.2.1 Architecture

We chose a web based approach for EMBERS visualizations for many reasons. Since EMBERS is already a distributed system spread over a cluster of virtual computing resources, the system already deals with web technologies. Server side tasks can add officially submitted forecasts to the web application immediately upon generation. Asynchronous tasks handle preprocessing steps needed for displaying forecasts such as converting place names to latitude and longitude for display on a map. Also, we were able to leverage many open source technologies for building web based visualization tools such as Django for web-server development and d3.js for data visualization [11]. Also, a web based architecture means that we can serve up simple JavaScript code to clients and relegate heavy computations to server-side or cloud-side. Being web based also gave us the ability to set up user logins so that we could have an idea of who accesses the site and their typical usage patterns.

While the web based architecture seemed best suited to our needs, it presented several challenges due in large part to the quantity of data EMBERS uses and produces. Not only will displaying all forecasts for a given date range at once likely be too much for a user to process, the time required to send the data for hundreds of forecasts from server to client can increase page load time to unusable levels. Infinite scrolling techniques solve this issue

for lists of forecasts by showing a reasonable number of forecasts at first and loading more from the server as scrolling reveals more. For aggregate visualizations of forecast dimensions, calculating values server side saves the need to spend time doing such calculations in the client's browser. Rather than sending all records for the client-side code to do aggregation, the server can do those calculations and pass only aggregate values to the client.

A web application has the benefit of allowing many users to easily view visualizations of the EMBERS system. This allows users to access views with browsers already installed on their machines. To keep user interaction at realtime speeds we were able to leverage a complex backend that uses asynchronous tasks to prepare and prepocess data before users make requests for it. This architecture not only avoids users waiting for computation to finish but also reduces the amount of data that needs to be send to the client's browser which reduces page load times.

2.2.2 Display

The forecast overview page was designed to give an at-a-glance overview of the state of the EMBERS system. To that end, the page displays several aggregate views of forecasts generated for a time period and a table listing the details of each of those forecasts. All the elements of the page are connected in such a way that interactions with one element are reflected in other elements. Using brushing and linking techniques [12], users can apply a filter on a chart representing one dimension and the interface will update all other charts to remove items that were filtered out by that interaction. This way, users can discover information relevant to whatever their goal may be.

Since the date on which events are predicted to take place is one of the more important features of a forecast, the top center of the display features a bar chart of the count of

forecasts per day, as shown in Figure 2.2. Daily resolution is appropriate here because forecasts include only predictions for the date of an event (not time of day). Users can drag on this chart to draw a box which defines a date range for filtering the display. This chart is linked with all other components of the page so all other charts and the table of forecasts update to include only forecasts with event dates in the given range. Once drawn, the selection box can be resized via handles on either side.

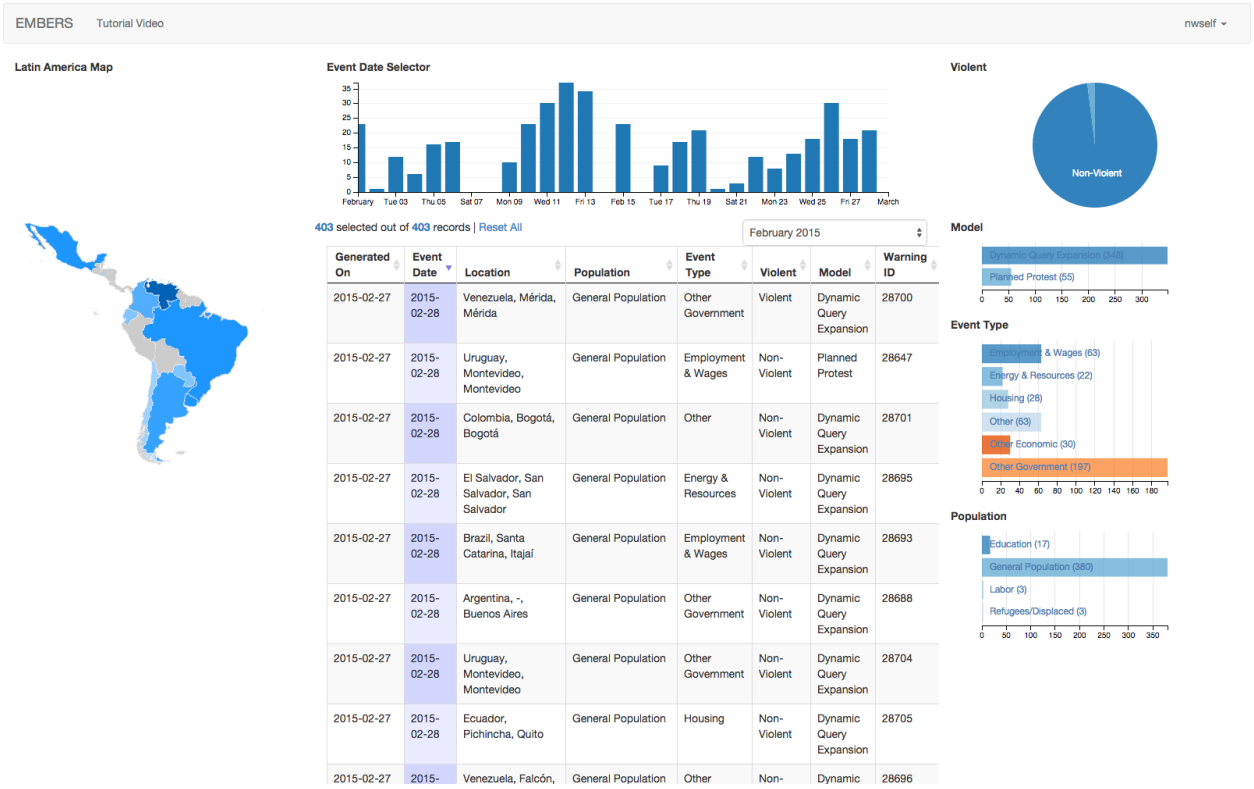


Figure 2.2: Overview page. Used with permission of Dr. N. Ramakrishnan, 2015.

The next focal point of the page is a table of forecasts underneath the date range selector. Since the goal of this page is to discover what forecasts have been made and possibly to inspect them more closely, this is a primary element of the page. The table lists all the relevant dimensions of the forecasts: the date on which the forecast was generated, the date on which an event is forecasted to take place, where the event will take place, which subgroup

of the population will take part, about what will they be protesting, whether the event is forecasted to be violent or nonviolent, the model that generated the forecast, and the ID of the event so that it can be referenced outside of the webpage. The table can be sorted by any of these columns by clicking on a column header label. Secondary sorting can be done by pressing the shift key and selecting any other column header label. Carets on the column headers indicate the current sorting direction. Selecting a row in the table opens up the audit trail visualizer in a new tab displaying that warning's audit trail (described later in Section 2.3).

Around the periphery of the date chart and forecast table are various charts that provide aggregate information about the currently selected set of forecasts. In this way, users can understand at a glance the broad nature of the forecasts in their filtered set without having to trudge through a long list of warnings making a mental tally. Along the right, there are horizontal bar charts for the counts of which model generated forecasts, counts of population type, and counts of event type. There is also a pie chart for the percentage of violent and nonviolent forecasts. Each of these charts affords the same interactions. Each bar or pie slice in one of these charts is a boolean OR (union) selector that adds forecasts with that value for the charts dimension to the filtered set. By default all selectors are active so no forecasts are filtered out by these charts. Selecting one of these elements takes the chart out of all-on mode and sets that bar or pie slice to active while deactivating all other values for that dimension. Other values can be added by selecting their element to create a query with any combination of these values. Since each of these dimensions are categorical and mutually exclusive within themselves, this allows all boolean queries on these dimensions to be expressed.

To the left, a choropleth map of our region of interest is displayed. Countries in the region are colored based on the number of warnings for that country compared to all other countries

in the region. Countries on the map behave the same way as the bar charts along the right. Selecting a country filters out all forecasts for all other countries and more countries can be added in by selecting more. In this way, users can geographically narrow their search and see the lay of forecasts across the region.

Throughout the page, pointer shape changes to indicate the ability to interact. For the date selector bar chart, when there is no selector box the mouse pointer becomes a crosshair to indicate that drawing is possible. When there is a selector box, the mouse pointer will change to resize and drag pointer shapes to indicate these interactions as appropriate. All of the boolean OR filter features (aggregate bars, pie slices, and country shapes) are highlighted on hover and the mouse changes to select shape. Each chart has a reset button that appears when that chart is currently being used for filtering. Also, above the central table of forecasts, there is a reset-all button that will change all charts to do no filtering. Near this reset-all button is a count of how many warnings are currently displayed (in the filter set) and how many total warnings there are in the current date range.

2.3 Audit Trail Visualizations

2.3.1 Architecture

The displays for audit trails build off of the architecture in place for the forecast overview page. Whereas the forecast overview page needed only a database of officially submitted forecasts, to visualize all the data that was a part of a forecast requires an entire audit trail. As a server-side processing step, whenever a new forecast submission is detected, a process is spun up to compile all the messages referenced by the forecast into one audit trail JSON object that contains the entire provenance of the forecast.

2.3.2 Display

Because the structure of an audit trail depends entirely on the model a different visualization is needed for each model that we aim to display. In order to give these displays a consistent visual feel, we developed a quadrant based approach. For the display of an audit trail, screen real estate is split into four quarters. The top left quadrant always displays the tree-like structure of the audit trail which depicts the processing of data from raw input all the way to submitted forecast. The top right quadrant displays the selected forecast in the context of other forecasts for events around the same time. The bottom two quadrants are reserved for model-specific visualizations of raw input and derived features.

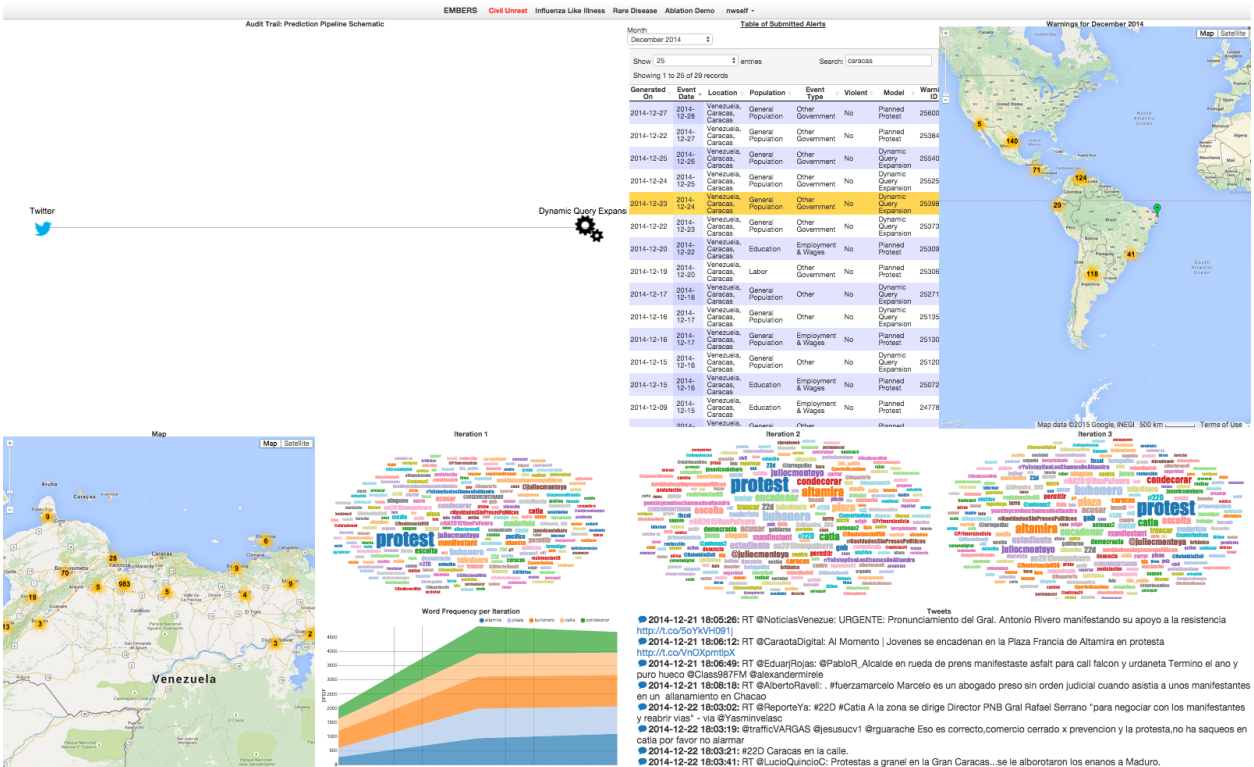


Figure 2.3: Quadrant view for a DQE forecast. EMBERS screenshot used with permission of Dr. N. Ramakrishnan, 2015. Maps generated with Google Maps Javascript API, used under fair use.

2.3.3 Top Quadrants

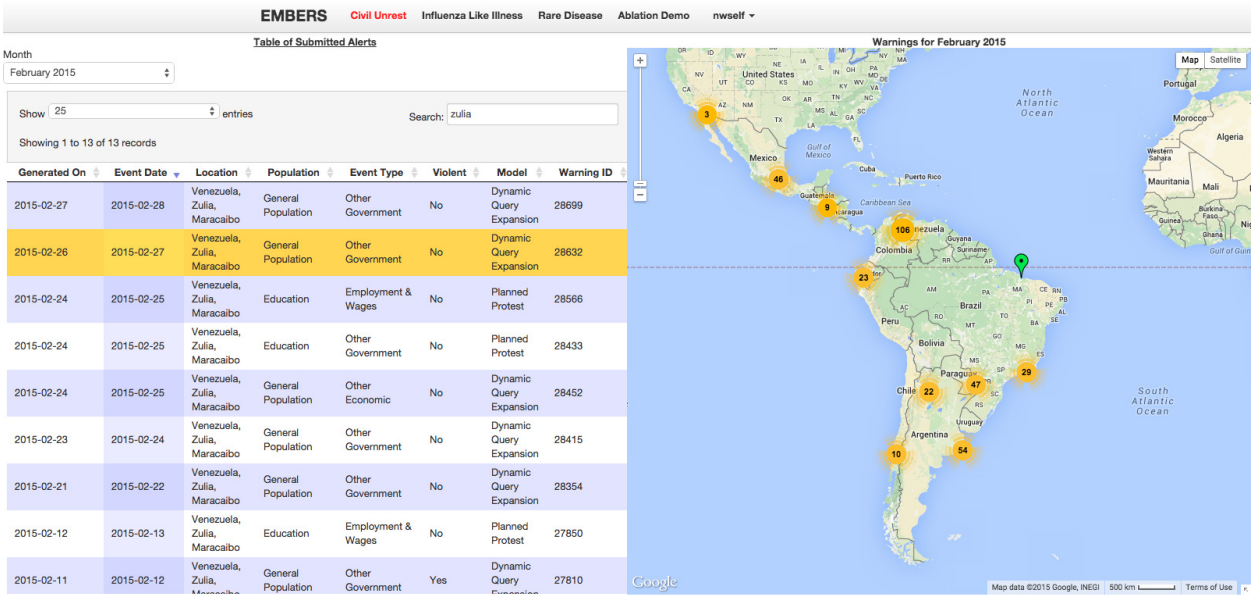


Figure 2.4: Top quadrant map and forecast list. EMBERS screenshot used with permission of Dr. N. Ramakrishnan, 2015. Maps generated with Google Maps Javascript API, used under fair use.

The top part of the display gives a general overview of how the currently selected forecast came to be and how it fits in with other forecasts in the region. To the left, a tree represents the generation of this forecast with nodes representing messages and edges representing processes that act upon the messages on their left to create the message on the right. Though we expect most of our intended audience to assume that this schematic is laid out from left to right, users can discover this by working backwards from the single final message on the right to the input data on the left. In some cases, the tree has to be simplified to avoid visual clutter. For instance, models that use tweets to make predictions often use a few hundred tweets. For the Spatial Scan model, which detects clusters of tweets over time and space, the tree simply displays clusters of tweets as inputs to the model node (as shown in Figure 2.5 rather than crowd the space with hundreds of tweet icons. Also, this schematic view hides the complexity of enrichment steps such as geolocation and natural language processing by

labeling the enriched messages and not displaying nodes for all the raw messages from which they were derived. To the right, the region-wide context of the given forecast is displayed with a table of forecasts for this time period and a map of the region with forecasts plotted on their predicted locations. The map uses Google's default marker image to mark locations of predictions but when many markers would overlap they are combined into a circular marker that indicates how many predictions there are in that area. Zooming in will cause these markers to recalculate and potentially split apart into many markers. Clicking on a group marker will zoom the map to the region covered by that marker and recalculate marker positions for that zoom level and area. This table of forecasts has the same functionality as the table of forecasts on the overview page and allows searching data in any column of the table via the search box.

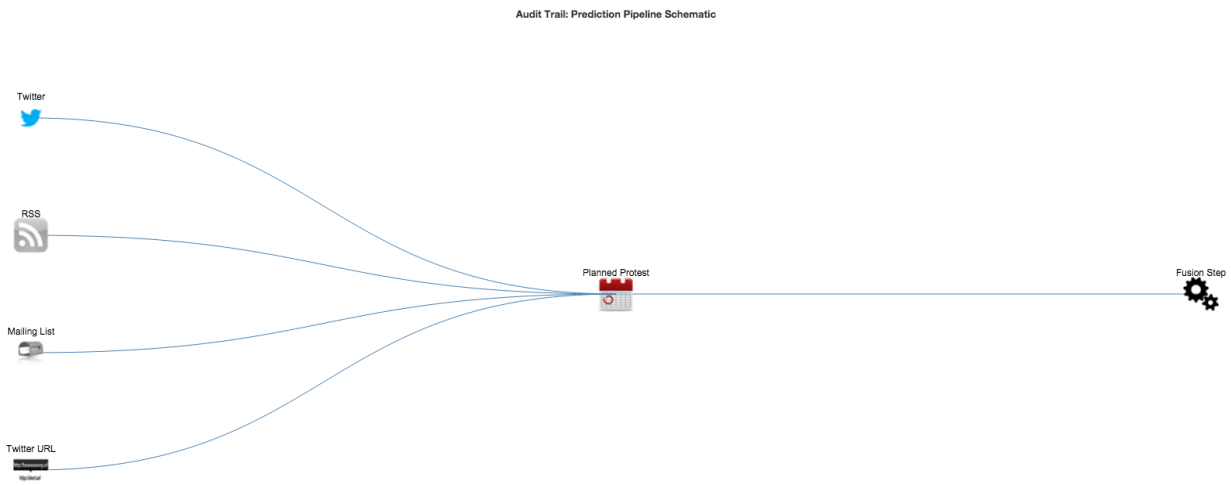


Figure 2.5: Top quadrant audit trail generation schematic. Used with permission of Dr. N. Ramakrishnan, 2015.

Table 2.1: Models by event class and input sources

Model	Event Class	Inputs
Planned Protest	Civil Unrest	Email, RSS, Twitter
Dynamic Query Expansion	Civil Unrest	Twitter
Spatial Scan	Civil Unrest	Twitter
LASSO	Civil Unrest	Blog, ICEWS, Inflation, RSS, Tor, Twitter
Bayesian Fusion	ILI	Google Flu Trends, Google Search Trends, Health Map, ILI Counts, Weather
Hantavirus	Rare Disease	RSS
Currency Deltas	Finance	Currency data, RSS

2.3.4 Bottom Quadrants

The bottom half of the screen is reserved for model-specific visualizations. Because each model uses different inputs, extracts different features, and uses them in different ways, each model has different visualizations in this area. What follows is a short description of these models and the visualizations that go with them.

Planned Protest

The planned protest model looks for trigger phrases in tweets and articles that indicate an event will occur [13]. If such a phrase is found, the model looks for a time phrase and if this phrase is resolved to be in the future then a forecast is generated. Since this is a relatively simple model, the visualization simply lists the trigger phrase and time phrase which were the determining factor in the forecast and displays the original content in which these phrases were found.

Figure 2.6: Planned protest with trigger words highlighted. Used with permission of Dr. N. Ramakrishnan, 2015.

Dynamic Query Expansion

The Dynamic Query Expansion model builds a vocabulary out of by querying tweets until a threshold is reached [14]. This model starts with a seed set of trigger terms that indicate civil unrest. It queries for tweets from a given day in a given country that contain those terms. Then the model adds terms that co-occur with the seed set to the seed set for another iteration of this process. The model continues to iterate until it reaches a stopping condition at which time it generates a forecast if necessary. To visualize this model, we chose to emphasize the idea of iterative improvement of the search set until the threshold for sending forecasts is met. To do this, a series of three word clouds indicates terms in the search set from the first, middle, and last iterations. Word size is proportional to each word's *document frequency-inverse document frequency* score which is a measure of a term's importance to a tweet offset by how often that term appears in any tweets [15]. The font color for a term is consistent between the word clouds for all iterations. In this way, users can browse a collection of words that are related in some way to this forecast to get a better idea of what it may be about. Furthermore, a stacked area chart displays the *df-idf* scores for the most important terms out of all the iterations which gives an idea of the spread and importance of these terms. To give geospatial context to this data, to the left there is a map of the country for which this forecast was generated with the locations of all the relevant tweets plotted. To avoid visual clutter, nearby markers are combined into cluster markers. This

gives the user an idea of where the issues involved in this forecast are being discussed. Users can also trace the forecast down to the level of individual tweets by browsing the list of all tweets that were used in this forecast. This way users can find out what social media users actually had to say about the terms identified by the model.

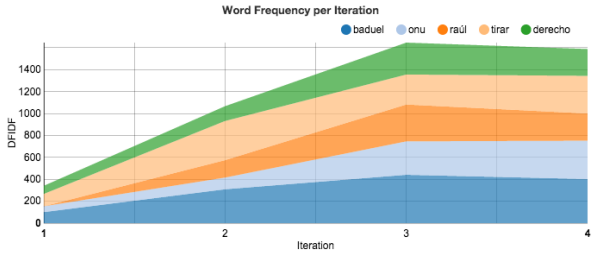


Figure 2.7: Stacked area chart for *df-idf* scores for top keywords over DQE iterations. Used with permission of Dr. N. Ramakrishnan, 2015.

Figure 2.8: Word clouds for selected DQE iterations. Used with permission of Dr. N. Ramakrishnan, 2015.

Spatial Scan

The spatial scan model monitors tweets that contain keywords from a dictionary of protest related vocabulary [16]. These tweets are tracked temporally and geospatially. If the intensity of protest-related tweets grows fast enough to cross a threshold, the spatial scan model generates a forecast. To visualize the geospatial aspect of this model, the tweets are displayed as clustered markers on a map of the region. For the temporal aspect, a stacked bar chart displays the growth of the most prominent keywords. Each bar represents tweets from one window of time that was an input to the spatial scan model. To get an overview of issues

involved in this event there is a word cloud of all the related protest keywords. To trace the forecast back to the source material, there is a listing of all the tweets used in a given forecast. In this listing, important keywords are highlighted in the color matching their color in the keyword word cloud. Furthermore, the temporal cluster in which this tweet belongs is indicated by a colored speech bubble at the start of the line. Since the time windows for temporal clusters can overlap, tweets belonging to multiple clusters have multiple speech bubbles.

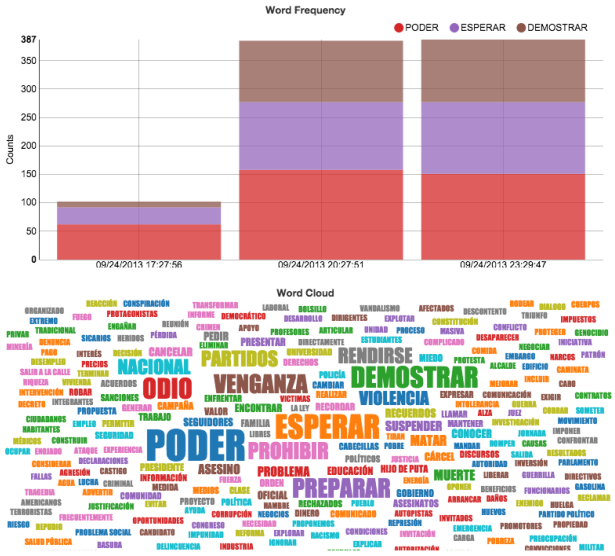


Figure 2.9: Stacked bar and word cloud for spatial scan. Used with permission of Dr. N. Ramakrishnan, 2015.

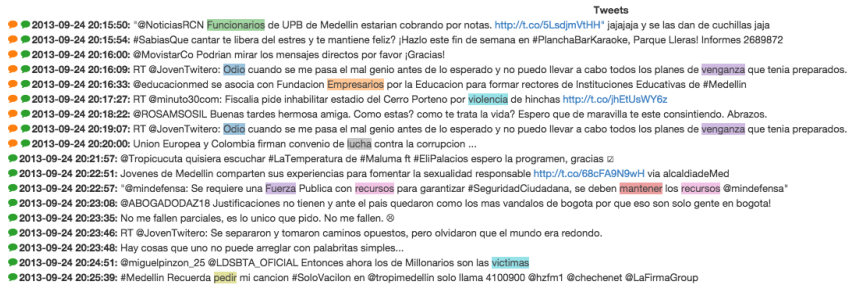


Figure 2.10: Tweets for spatial scan with overlapping cluster speech bubbles. Used with permission of Dr. N. Ramakrishnan, 2015.

Influenza Like Illness

Like the LASSO model, the influenza like illness (ILI) predictions synthesize data from many time series datasets [17]. Unlike the LASSO model which forecasts discrete events, the goal of ILI predictions is to forecast the total count of reports of influenza like illness in a country for a month. The predicted counts for each country appear in the table of all submitted predictions for the given month. The visualization for each of these forecasts takes largely the same form as the LASSO model's with important time series datasets displayed in the detail quadrants of the display. In this way, a user can get a feel for the trends in interesting datasets such as ILI counts from previous years, temperature and humidity data, and trends in Google searches at the time leading up to the prediction.

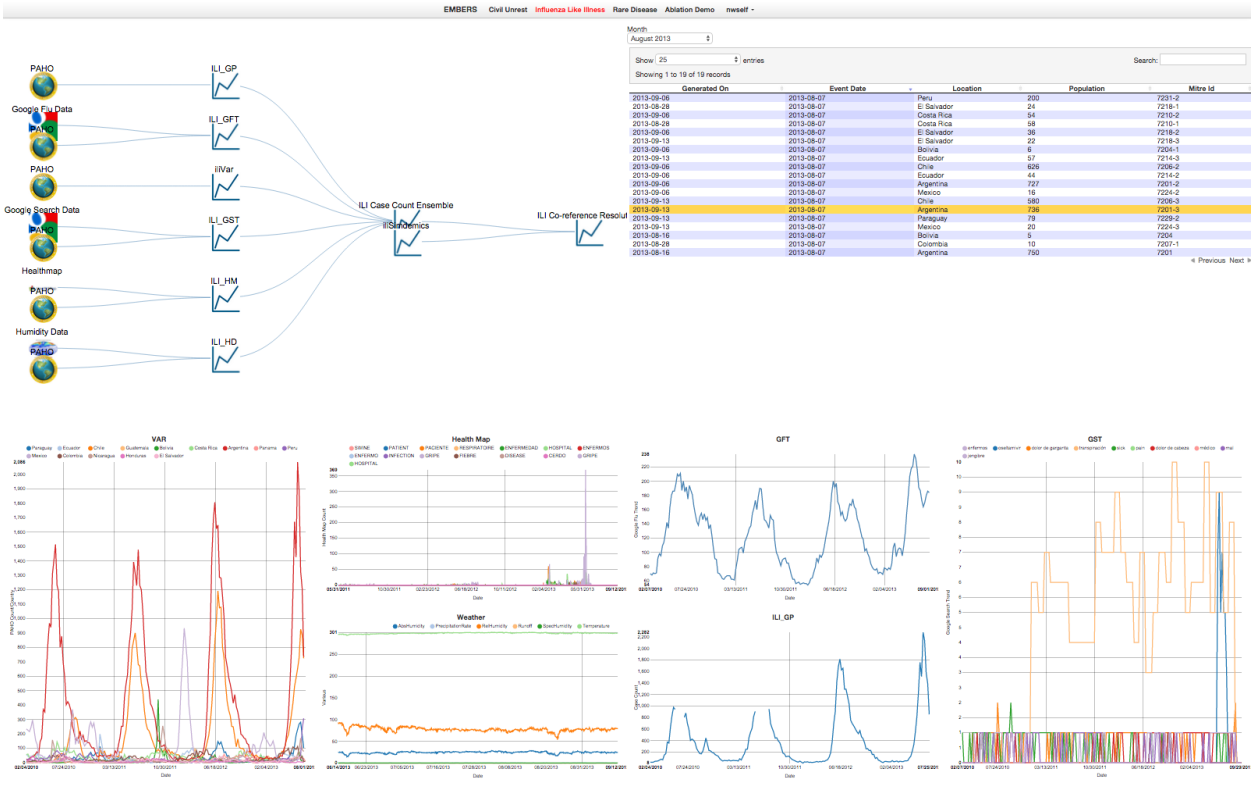


Figure 2.11: Influenza like illness audit trail view. Used with permission of Dr. N. Ramakrishnan, 2015.

LASSO

The LASSO model performs logistic regression using the least absolute shrinkage and selection operator (LASSO) method to predict the probability of an event happening in the future [18]. This model can work on features derived from many different sources. To extract features from text data, a dictionary of country-specific keywords was curated by a group of subject matter experts. The different input sources used by the LASSO model include counts of events identified by the Integrated Conflict Early Warning System (ICEWS), the count of daily users of the Tor network for anonymous online activity, the value of the country's currency relative to the US dollar, and counts of the occurrence of protest-related keywords per day in tweets, blogs and news articles. Since this model synthesizes many time series datasets, we visualized this with time series plots where appropriate. In most cases, the features extracted from each data source are displayed. Because the ICEWS data takes the form of many discrete entries and since many different features are extracted from it, we display this information in tabular form.

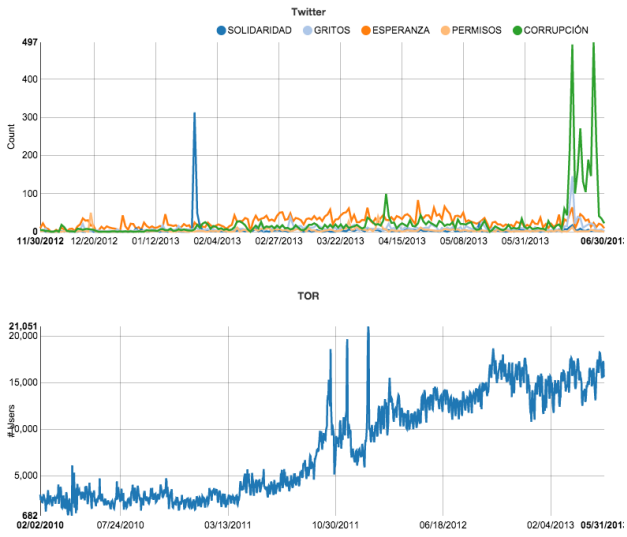


Figure 2.12: LASSO detailed view for Twitter and Tor data. Used with permission of Dr. N. Ramakrishnan, 2015.

2.4 Ablation Demo

The visualizations for other models were designed to summarize the data that led to a forecast's generation. These displays allowed users to answer questions about forecasts like what is going to happen, what part of the population will protest, and why.

2.4.1 Requirements Analysis

The displays for summarizing audit trail data do not support more complex hypothetical questions which can be especially useful for sense-making [19]. The multi-source nature of the LASSO model allows us to frame a what-if question that asks what if we do not include some data sources. Would LASSO still generate the same predictions if we exclude a set of inputs? We call this process ablation, i.e., removing one part from a whole while leaving the remaining parts intact.

Another motivation for this type of interface is to allow users to improve their trust in EMBERS's forecasts. Because the other civil unrest models covered so far handle only one input stream, the issue of believing in the value of the data source reduced to the problem of believing in the value of that one data source. For instance, to evaluate a Planned Protest forecast, if a user does not agree that the combination of the identified trigger phrase and time phrase from the source data does not actually signal that an event will take place, then that user will not trust the prediction. Evaluating the output of the LASSO model is more complicated in part because reading all the input data would be a time-consuming, tedious task and even still trends in the features would not be clear. Because LASSO works on many sources, if an analyst does not trust a particular input stream, it is possible to rerun the model with all the inputs except the untrusted one.

2.4.2 Design and Implementation

We designed an interface for users to ask the question of what if we remove a set of input data sources from the LASSO model. This interface was meant to be a demonstration and prototyped on month's work of EMBERS forecasts (and data). The initial view displays all the forecasts generated by LASSO with the benefit of all input sources. As users remove data sources from consideration, the visualization displays a new set of forecasts for the month that leverages only the data sources still considered.

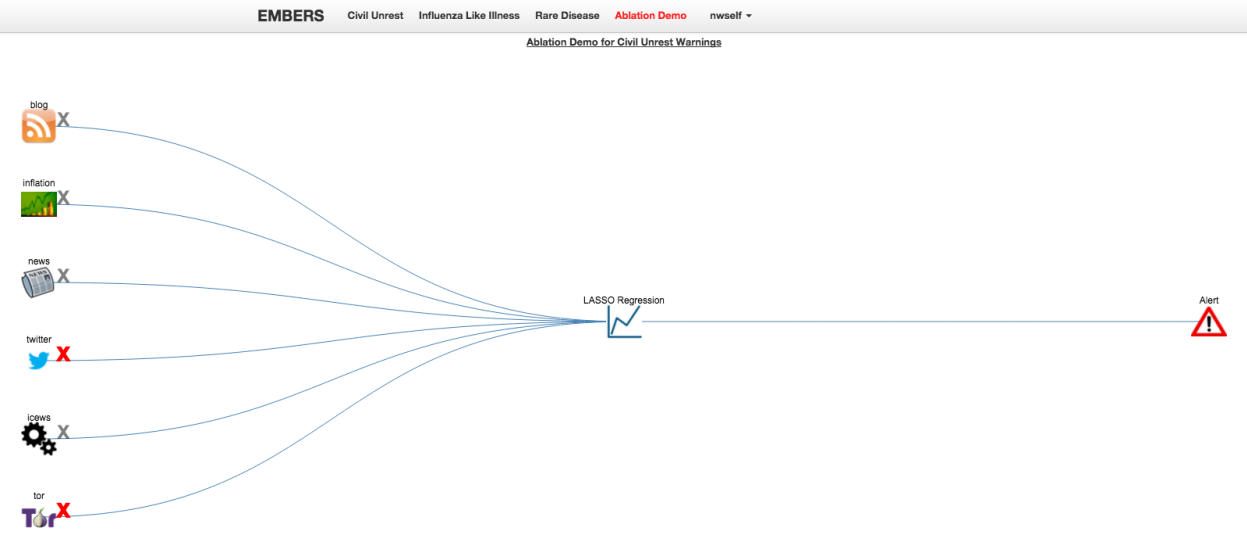


Figure 2.13: Ablation schematic view with Twitter and Tor ablated. Used with permission of Dr. N. Ramakrishnan, 2015.

For the design, we leveraged a similar layout to the audit trail views. The forecast generation message schematic is the primary means of interaction. This tree shows a generalized interpretation of LASSO forecast generation. Each of the data sources has a node which points to the LASSO models node. Each of these data source nodes have buttons which will ablate, or remove, those data sources as inputs. Next to the schematic is a table of forecasts generated by the LASSO model with the current set of input sources and a calendar displaying the distribution of those forecasts over the month. This table and the associated

count of forecasts in the table allow a user to answer questions such as will EMBERS still make the same forecasts if we exclude some data. So, for instance, if a user removed from consideration all the input sources except Twitter, then the table will update to show the set of warnings that LASSO generates for this month when it considers only Twitter and the calendar will update to show the days for which those events are forecasted.

The remaining quadrants are left for detailed views of each data source. In most cases this takes the form of a time series. When a user removes a source, the chart for that source is also removed and replaced with a note reminding the user that this source has been ablated. Furthermore, during its training period, the LASSO model assigns an importance to the signal from each source and if that value is too low then that source is not used in forecast generation. These weights change with each combination of input sources, but in cases where the importance of a source is too low, the chart for that source is replaced with a note informing the user that this source is unused in the current configuration. In this way, users have a view of which data sources are relevant to a given configuration and what the historical trend of features relevant to forecast generation is for that configuration.

2.5 Evaluation

This interface was used as a visual aid for demonstrations of the EMBERS project to project stakeholders, funding agents, and other interested parties. If anyone was interested in the site, they could request an account through the website. This account gave them access to all the visualizations for forecasts and audit trail data for approximately 33,000 forecasts generated by EMBERS. Account holders could try out the site and give feedback through several channels. With this feedback, we went through several stages of iterative design, refining the details of the interface. After three years of the EMBERS project, there were 166

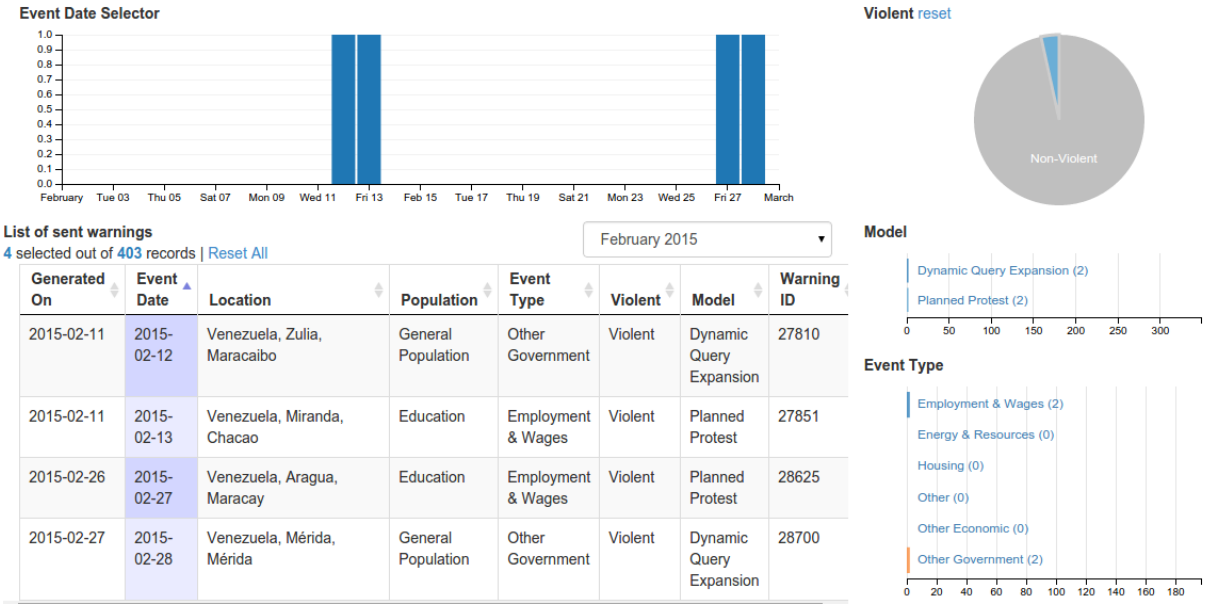
registered users of the site. The next section gives an example of a possible workflow through the site to determine what forecasts to examine and why those forecasts were submitted.

2.5.1 Case Study

To demonstrate the usefulness of these visualizations we provide a case study. The data for the case study is comprised of predictions generated by EMBERS from June 2013 through April 2015. This includes prediction information and the audit trails for those predictions. The task is to do unstructured exploration of the data to investigate civil unrest in Venezuela. The intended user for this type of exercise is an intelligence analyst who aims to investigate protests and contexts surrounding them in Venezuela.

On a hunch, the user decides to investigate February 2013. To do this, she changes the dropdown menu for the displayed months to February. This updates the site to show all predictions for that month which updates the table of predictions and all the aggregate charts. She notices that there are 403 predictions for this month across Central and South America. On the choropleth map, Venezuela is filled in with a darker blue than any of the other countries, indicating that there are more predictions for that country in this month than any other country. By hovering over Venezuela, the user can tell from the hover text that there are 116 predictions for Venezuela this month. In contrast, Brazil's hover text indicates that there are 49 predictions for this month, so Brazil is lighter blue. The user clicks on Venezuela to filter the display to show only predictions for this month that are for Venezuela. This changes all the aggregate charts to show only the predictions that are for Venezuela. The website animates this change so that relative changes between the previous and subsequent views are easier to follow.

The user notices that there is a small area of the pie chart that indicates that some of the



Generated On	Event Date	Location	Population	Event Type	Violent	Model	Warning ID
2015-02-11	2015-02-12	Venezuela, Zulia, Maracaibo	General Population	Other Government	Violent	Dynamic Query Expansion	27810
2015-02-11	2015-02-13	Venezuela, Miranda, Chacao	Education	Employment & Wages	Violent	Planned Protest	27851
2015-02-26	2015-02-27	Venezuela, Aragua, Maracay	Education	Employment & Wages	Violent	Planned Protest	28625
2015-02-27	2015-02-28	Venezuela, Mérida, Mérida	General Population	Other Government	Violent	Dynamic Query Expansion	28700

Figure 2.14: Summary page for predictions for violent unrest in Venezuela for February, 2015. Used with permission of Dr. N. Ramakrishnan, 2015.

predictions are for violent protests. She clicks on this pie chart area to filter the display to show only predictions for violent protests, as shown in Figure 2.14. There are only four warnings that fulfill the criteria of being in Venezuela and being for a violent protest and the website now displays those. At this point, the event date selector shows that there are violent predictions for weekends at the middle and end of February and the table lists these four predictions. The user decides to investigate the audit trails of these predictions to decide if they have enough merit to warrant preemptive action for safety. She wants to investigate them in ascending chronological order so she clicks on the column header for Event Date to sort them in that order.

She clicks on the first one, which is a Dynamic Query Expansion prediction, and this opens up a new tab with a visualization of the audit trail for that prediction, as shown in Figure 2.15. She first checks out the stacked area chart, as shown in Figure of the top keywords and finds that they seem to point to unrests involving injured (*herir*) students (*estudiante*) in the

Venezuelan state of Táchira, and the next day February 12 (*12F*). The word cloud for the final iteration of the DQE algorithm, which indicates the most important keywords during the iteration which sent out the prediction, includes protest related terms (*marcha*, *protest*, *manifestant*) and relevant hashtags that call for marches on February 12 (*#YoMarchoEl12F*). This prediction seems to have to do with a march to protest some injury to a student. To get a more precise understanding of the situation, the user reads some of the raw tweets that the DQE model found to be relevant and discovers tweets that general discontent with lack of transparency in the government is causing a great deal of unrest.

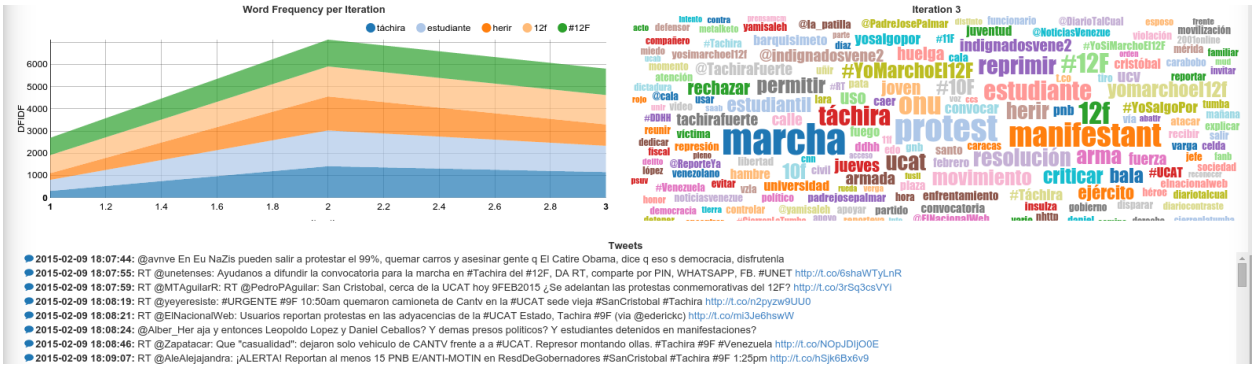


Figure 2.15: Audit trail view for a prediction for violent protest in Venezuela in February, 2015. Used with permission of Dr. N. Ramakrishnan, 2015.

With some idea of what this prediction is about, the user switches back to the summary page and opens up the next prediction which is a Planned Protest prediction for the next day. This prediction is based on a tweet that contains the words protest (*protestar*) and tomorrow (*mañana*). This tweet does not explicitly call for protesting, instead it says "do not protest, instead spend time teaching your children to wait in line and beg for food." Without context it is hard to tell if this is sarcasm. Regardless the user is suspicious of this event because it did not clearly schedule a protest.

Before investigating the audit trails for the final two violent predictions for February, 2015, the user decides to see if there are any interesting patterns in the data. She clicks the reset

button for the violence chart to return the page to displaying all the predictions for Venezuela and notices a spike in predictions on February 13, as shown in Figure 2.14. To investigate the predictions for that day, she drags a selection box on the event date selector graph to select only February 13. This filters the rest of the page to show only the 16 predictions for that day in Venezuela. By browsing the table of warnings, the user notices there are several predictions with event type of Employment & Wages predicted by the Planned Protest model. By inspecting the audit trails for each of these five predictions, she notices that they each were triggered by retweets of the possibly sarcastic tweet that triggered a prediction for violence. By noticing that the locations for these five Employment & Wages predictions include locations in various parts of the country, she comes to the conclusion that there is discontent across Venezuela due to economic problems. And in fact, from the end of 2014 into 2015 Venezuela experienced a series of protest and civil unrest due to food shortages, violence, corruption, and a failing economy.

2.5.2 Conclusion

Because we used the site for telling many stories similar to the case study during presentations, we believe that the interface was successful at allowing exploration of forecasts and audit trails. These presentations typically covered the what and why of an event. Additionally, feedback from users about desired functionality implies that they were able to use the site enough to form opinions about enhancements. As described in Section 2.4, it is possible to support complex what-if questions, but the ablation interface that we prototyped has the limitation of only allowing for questions that involve removing data sources.

Figure 2.16: Summary page for predictions for unrest in Venezuela for February 13, 2015. Used with permission of Dr. N. Ramakrishnan, 2015.

Chapter 3

Reverse OSI

Data mining problems are often framed in a signal-to-noise ratio metaphor. EMBERS aims to detect signal from the data streams that it ingests. In particular, one of the tasks of EMBERS is looking for signal that serves as a precursor to civil unrest events. When the signal-to-noise ratio is low, there is not much evidence that anything is happening; this can make event forecasting difficult.

3.1 Requirements Analysis

The goal of the Reverse OSI methodology is to enable analysts to rapidly work backward from specific GSR (ground truth) entries to identify indicators that suggest that such an event will occur, or was likely to happen. In other words, the goal is to use the benefit of hindsight to identify precursors. These indicators need not be numeric or easily quantifiable at this point, but rather human and impressionistic. Some of these indicators might actually be specific causes (i.e., factors that make an event occur). In this case, we need to record that they are actually causes (this may make a difference in later analysis that looks at what

other indicators we can think of that would be generated by these same causes).

3.1.1 Methodology

The Reverse OSI methodology aims to break down each GSR entry into four aspects:

1. Context/Contextual Factors
2. Trigger Event
3. Whether the protest is spontaneous or organized?
4. If organized, who are the political entrepreneur(s)?

These four aspects are explained in further detail below. Questions are given at the end to help elucidate the information necessary for the above four aspects.

Context/Contextual factors are those events and considerations that form the backdrop of the event. These should not be too general, but specific to the type of event (economic, resource, etc.) and the characteristics of the country (e.g., the issue of justice for victims of government repression in the fight against guerrillas is still an issue in Colombia, Peru, and Mexico but no longer in Argentina; export taxes are an issue in Argentina but not in Chile). Context tells us who is in the "in-group" versus the "out-group" (e.g., use of the phrase 'neo-liberal' indicates a division between those favoring government control of a market and those favoring private enterprise and freer markets). Context can be provided at different granularities: (i) for an individual event (ii) at the level of a set of related, cascading, events (e.g., the Brazil protests in June 2013), or (iii) events related by a "cause" (e.g., coffee farmer strikes, protests against a tax law). The difference between (ii) and (iii) is that events in type (iii) can be quite distantly separated in time. For instance, coffee farmer strikes can

be separated by 6 months but they might be protesting for the same reason. Whereas in (ii) events are close by in time, and often are developments of one another. There is not a predefined list of what constitutes context, but rather this should be a matter of identifying what was relevant or special in the context for the event in question.

Triggers/Trigger events refer to an action committed by the government (e.g., passing legislation; police brutality, etc.) or any third party (e.g., criminal gang activity) or to a natural event causing human suffering (e.g., severe hurricane, major earthquake, etc.); these events are not produced merely by people participating in the civil disruption that we are trying to explain, but occur prior to the civil disruption and may or may not be causally connected to the disruption. Please identify any triggers for the considered events. Also, please identify any noticeable candidate triggers that did not lead to an event, i.e., things that look like they could have been triggers, but weren't. Do not spend time looking for candidate triggers that did not act as triggers, but if you do see them, please note them. Triggers that lead to events are usually followed by Political Entrepreneurs' actions, but in the rare case of spontaneous events they are not preceded by Political Entrepreneur activity.

Political Entrepreneur is someone who articulates a call for action in a manner that resonates with those who will participate in the event detected by the GSR. Lots of people will be making calls for action, but a Political Entrepreneur is someone who has a following. Political Entrepreneurs are not all-powerful; there must be a context within which action is perceived to constitute an appropriate means of expression by those who would participate in the event. Trigger events are likely to be necessary. In some cases, there might not be a "true" single political entrepreneur, in which case, please identify any characteristics of heavily involved populations or organizations, even if no one is acting as a true Political

Entrepreneur.

Spontaneous Events: Such events still have a context and triggering event, but communication regarding when and where to assemble gets its legitimacy from the triggering event, not from the sender of the messages (i.e., the political entrepreneurs). Spontaneous events are likely to be short-lived unless a political entrepreneur articulates and disseminates a vision in which continuation of the event is perceived to keep pressure upon the government or the authorities.

3.2 Design and Implementation

We developed a web interface for investigating events from the GSR and for registering comments about those events, with a view to elucidating the above four aspects. This website allows structured commenting on any event in an effort to support discussion among users about those events and what their causes might have been. With such a tool in place, we can easily distribute the task of investigating events across our team of researchers. In fact, although not investigated here, the Reverse OSI tool we built opens up the possibility of using crowd sourcing platforms like Mechanical Turk. Though it would be possible to use a machine learning approach for taking a set of events and looking backwards in time for signals that precede them, there are several advantages to a human approach. Humans will be able to apply cultural and social cues to comprehend an event and consider where precursors may be found. Plus, with a crowd sourcing tool, different users will be able to see the inputs from other users and have conversations. There is an added bonus to having this kind of infrastructure set up. We can investigate any questions we might have about the events that we are predicting. If we have any questions about the type of events we

are predicting we can add those questions to the commenting area for each event to compile feedback about them.

	Mitre Id	Headline	Date	Source	Country	State	City	Event Type	Violence	Population	MLE	PP	Size	LF
	10585	70 familias protestaron por retraso en entrega de viviendas	Aug. 30, 2013	El Nacional	Venezuela	Falcón	-	Housing	Non-Violent	Refugees/Displaced	No	No	400	8
	10584	Protesta de motorizados tranca el paso en la Francisco de Miranda	Aug. 30, 2013	Ultimas Notícias	Venezuela	Caracas	Caracas	Other Economic	Non-Violent	General Population	No	No	<100	309
	10583	Protesta de trabajadores tranca acceso único a la UCV	Aug. 29, 2013	Ultimas Notícias	Venezuela	Caracas	Caracas	Employment & Wages	Non-Violent	Labor	Yes	No	<100	309
	10582	Militantes chavistas rechazan en Caracas la intervención en Siria	Aug. 29, 2013	El Universal	Venezuela	Caracas	Caracas	Other	Non-Violent	General Population	Yes	No	<100	309
	10581	Trancan autopista en Acarigua en rechazo al subsidio agrícola	Aug. 28, 2013	Ultimas Notícias	Venezuela	Portuguesa	Acarigua	Other Economic	Non-Violent	Agricultural	No	No	100	4
	10580	Trabajadores denuncian desabastecimiento en Pdval	Aug. 28, 2013	El Universal	Venezuela	Caracas	Caracas	Employment & Wages	Non-Violent	Labor	Yes	No	<100	309
	10579	Protesta por la asignación de casas trancó la zona rural de Anzoátegui	Aug. 28, 2013	El Universal	Venezuela	Anzoátegui	Puerto la Cruz	Housing	Non-Violent	General Population	No	No	<100	16
	10578	Trabajadores petroleros tercerizados exigen cargos	Aug. 27, 2013	El Nacional	Venezuela	Zulia	Maracaibo	Employment & Wages	Non-Violent	Labor	No	No	<100	32
	10577	Denuncian sabotaje en recolección de	Aug. 27,	El Nacional	Venezuela	Miranda	Petare	Other Government	Non-Violent	General Population	No	No	<100	10

Figure 3.1: Beginning view of the Reverse OSI website. Used with permission of Dr. N. Ramakrishnan, 2015.

The Reverse OSI website was added on to the existing web application framework that runs the warning and audit trail visualizers. The main part of the page is taken up by the table of GSR events as shown in Figure 3.1. For each item in the table we list many of the fields coded in the GSR. Since each GSR event maps to at least one news article we can use the headline from that article as a meaningful caption for each event. Other data from the GSR that is useful for distinguishing whether an event is interesting to a user includes the date of the event, the source that published the article, the location of the event, and the population, event type, and violence codes submitted by the GSR coders. To these fields we add a few derived items that might help users find events that could use comments. Many of these

fields were added because they not only give users some indication of interestingness but also because they will help support further analysis after crowd sourcing comments.

Derived Fields

In determining the interestingness of an event, there are factors to consider other than those encoded in the GSR. The GSR simply lists that an event has happened but says nothing about whether that event is unique or well-attended or even whether EMBERS actually predicted it. We add some columns to the table for data that is not reported by MITRE in the GSR but which we compute to give more information to the user.

To help users decide if an event will be interesting to them, some columns display whether EMBERS predicted an event. The Planned Protest column indicates whether the event was caught by the Planned Protest model. This model is fairly straightforward in that it simply discovers messages that announce an event in the future and sends a prediction about that event. The MLE column indicates whether an event was predicted by our baseline model which we call MLE. This model predicts events based on only the historical number of events for a location. To some extent this means that there is nothing particularly interesting about those events. We predicted them only on the basis that something usually happens in this location. In fact, in some countries, there are events every day such that simply predicting an event for each day performs reasonably well. By choosing a crowd sourcing approach for this project we have the ability to find out more information about the nature of the predictions and the events that EMBERS is looking for. We can test a hypothesis that events that the MLE model predicts are so commonplace that there is only marginal utility in predicting them at all. Users can comment on events that have been predicted by MLE and have the discussion about whether events predicted by the MLE model are worth the effort of predicting.

To give an idea of the impact of an event, we add a size field. Since size is not an official field in the GSR we estimate these numbers by running the GSR article text through some natural language processing and look for words that might indicate magnitude. With this information, users can prioritize events with large size or can debate the worthiness of predicting events with small sizes.

Location frequency is a count of how many events there were for the specific location of this event. With this information, users can find events that have small location frequency that were missed. Our hypothesis is that such events might be difficult to predict because they have a weak signal. Identifying clearer sources that might serve as precursors to such an event might identify a class of sources that is good for identifying other events from unique locations.

For each of these derived data fields we have a set of filters that can filter out the uninteresting events for that field. For MLE and Planned Protest columns, users can filter out any events that those models predicted. For Size users can filter out any events that had fewer than 100 participants. The filter for Location Frequency shows only events that happened at locations that have had no other events in the history of the GSR.

In addition to these filters, each row has indicators to show its status. A green smiley face indicates that an event has some comments while a red frowny face indicates events that do not. Users can favorite events to find them easily later by marking a star on a row.

For each of these types of user contribution there is a table level filter. Users can filter to see only events that they themselves commented on, events that anyone at all has commented on, or events that no one has commented on. Likewise, users can filter out events to see only those they have marked as a favorite.

Commenting

To get more information about or comment on an event, users click on the row for a given event to expand the row to show the commenting panel as shown in Figure 3.2. The expanded panel contains the rest of the information about the event so that users can make comments. This information includes the description from the GSR entry and links to the original article and any secondary sources added by GSR coders. There are several comment blocks to steer users towards the kind of answers that could be used to improve model performance.

The context comment is meant to provide a short blurb about why there is unrest. This is helpful not only as a short summary of why an event has happened but also as an English language description of the event which are almost always reported in languages other than English. This is a place for general comments on the event and on any previous chain of related events, if applicable.

Social scientists believe that unrest events are often sparked by some prior event. These trigger events could be a valuable source of information for prediction algorithms that might chain together events of different types to build a big picture of a situation and its many events. Examples of trigger events include new, unpopular legislation, imprisonment of a political leader, election of an unpopular leader, etc. For some events, there may be no clear, discernible trigger event at all. The trigger event comment is meant to support conversation about any type of prior engagement that might have sparked the unrest and is geared toward finding a signal that might have predicted this event.

The GSR event specification does not include whether an event was spontaneous or organized. Some social scientists believe that truly spontaneous events do not exist. To support answering this question, we added an dropdown menu to force choosing between only those two options. We also added a history section to show the history of changes between spon-

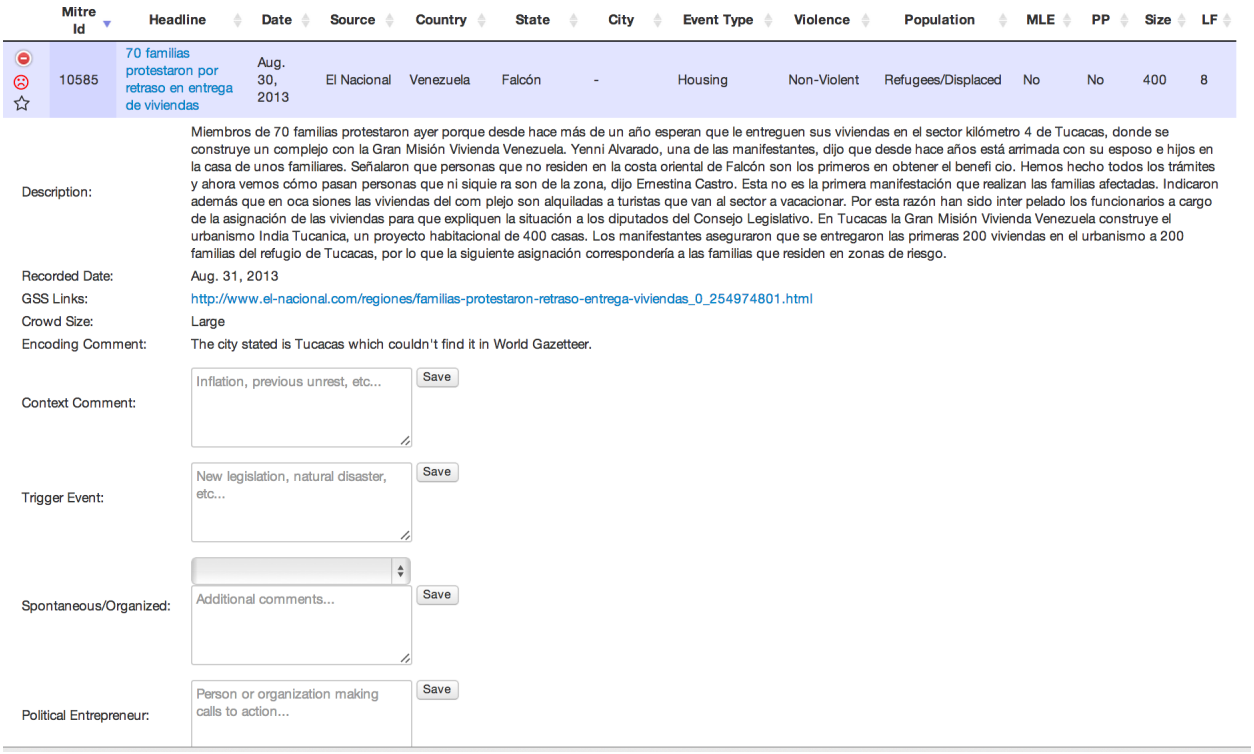


Figure 3.2: Reverse OSI website with an event expanded. Used with permission of Dr. N. Ramakrishnan, 2015.

taneous and organized and the user that made that change to give an idea of the discussion going on about this aspect of the event.

Since there might be public figures at the epicenter of unrest we add a political entrepreneur comment which is meant to describe any political actors responsible for making a given event happen. Though the OSI proposal does not allow tracking of individuals, it was not clear if this stipulation carried over to public figures. Regardless, results from this field could be relevant in the discussion of whether following public, political figures could be useful in attempts to predict civil unrest.

Finally, there is a comment field for any sources other than the sources reported by the GSR entry that describe this event. This field could point out news agencies that are not yet being monitored.

3.3 Evaluation

To evaluate this interface, we loaded it with all available GSR events. We asked users to inspect as many events as they cared to and fill out each type of comment for events that they researched. A list of several brainstorming questions was provided to users to give them an idea of what information to look for and include in their comments. For more information about these questions, see Section 3.3.1. The participants included social scientists and computer scientists from Virginia Tech and University of California, San Diego, that were members of the EMBERS research group. Over 7 months, these users contributed 2597 comments on 568 GSR events.

3.3.1 Brainstorming Questions

Analysts were provided the following guidelines for each aspect of the Reverse OSI methodology.

Context

Questions to consider include, but are not limited to:

- What was happening in the news relating to the subject of the unrest and involved population in the time leading up to the event?
- What was happening that may have contributed to the event but that was not directly covered in the news at the time? (Note that a good source for this may be retrospective news reporting on the event.)
- What was happening in peoples informal communications about the topic of the event

in the time leading up to the protest? (Twitter, blogs, etc.)

- How do any of the above answers differ from what we expect to see normally? (Note: It makes sense that identifying ways to measure and quantify these differences, and test that they are really differences as measured, will come from the technical side; the question here is to identify what is likely, e.g., "there was a lot of talk about wages, but there always is; this doesnt seem different from the usual," or "there is always talk about wages, but this was much more emotional and seemed to crowd out other common topics.")

- For a set of related events, also consider:

  - How closely do the events follow each other in time? Is there consistency?

  - Are there special relationships between the regions that engage close in time (the first set and the second, the second set and the third, etc.)?

  - Are there characteristics in common between regions that engage at the same time or close in time? Are these characteristics that are not shared by other regions?

  - Were there events that affected regions that engaged more than others, or that linked some of the regions that engaged to each other?

  - Which of the regions that engaged are regions that tend to have a lot of civil unrest anyway, and which ones are surprising locations? For those that were surprising, do they have any characteristics or events that particularly strongly linked them either to the topic or to the other locations involved?

  - Did the topic shift over the course of the events? What was going on as that happened, both in external events and in the mood and discussion of people participating in the events and reacting to them? How did this seem to occur: Are the topics related in nature, are they of interest to the same populations,

were they simply other topics that had a lot of restlessness around them and populations who were ready to join protests as they occurred, etc.?

  - What kinds of events and discussion were happening as the events continued and grew? What kinds of events and discussion were happening as they decreased and stopped?

- Are there consistent and predictable aspects of an OSI category, such as certain organizations who are active (even if they are not political entrepreneurs), certain topics, certain time frames, etc? (For example, are there consistently education-related protests in given locations at certain points during the academic year?) Are any of these matters consistent in describable sets of protests that do not match the OSI categories (perhaps, for instance, protests by urban populations in a given country)? This could be for a given location, in either case.

Trigger

- Was there a specific trigger event that led to the protest or other event? What was it? How do you know it was the trigger?

- What was happening in the news relating to the subject of the unrest and involved population in the time leading up to the event?

- What was happening that may have contributed to the event but that was not directly covered in the news at the time?

- For a set of related events, also consider:

  - Do the events seem to trigger each other, or do the events seem caused by a

common trigger, or some mixture or alternative? What contributes to this impression?

  - What kinds of events were happening as the events continued and grew? What kinds of events were happening as they decreased and stopped? Was there a trigger event for cessation, or did it seem to happen organically, or was there an identifiable but gradual shift? How much of this came from specific events and responses (e.g., demands were or were not met), and how much of it was more a matter of mood?

- Are there known environmental factors that would cause us to look for specific kinds of triggers for a category of OSI events, or for a definable type of event that does not match an OSI category? This may be within a given location in either case.

Political Entrepreneurs

- Was there a specific, identifiable Political Entrepreneur for this event? Who was it? Was there more than one? Who were they? Were they people or organizations? How can you tell they have a following? What was the reaction to their call for the event?

- Were there key organizations or organizing individuals involved in making the event occur? What were they saying and discussing – both in content and emotional tone in the time leading up to the event, besides any specific announcement of the event itself?

- How does any of the above differ from what we expect to see normally?

- For sets of related events, were there any changes in key organizations or key individual players involved? Did some join after the events started, did some become more or

less prominent during the series of events, etc.? The same kinds of questions about characteristics and linkages apply here as for regions, in the context category.

- Are there consistent and predictable political entrepreneurs for an OSI category of events or a definable type of event that does not match an OSI category? This may be within a given location in either case.

3.3.2 Worked out Example

We provide an example of a Reverse OSI analysis to illustrate the output of the methodology.

GSR Event

On August 8, 2013 300 businessmen engaged in a public demonstration in the south of Bogota. They claimed that their sales had fallen by up to 60% because of the citys increasing restrictions on alcohol sales.

3.3.3 Context

Bogota Mayor Antanas Mockus decrees "Carrot Law." In Colombia, 'Carrot' means someone who neither drinks nor smokes. According to the law, the sale of liquor was restricted to before 1 am. Liquor restrictions are traditional in Colombia on election days, when the sale of liquor is prohibited from 6 am of the day before to 6 am of the day following. Mockus decreed the control of liquor sales at the same time he imposed gun restrictions as a way of dealing with the high levels of violence in Bogota. In 2002 the Carrot Law was rescinded, with bars and clubs allowed to remain open until 3 am; this remained in place until 2011.

In 2008 the press began to report increasing concern about citizen security, and raised the

possibility of the Carrot Law being revived in targeted neighborhoods of Bogota. In January 2009 Councilwoman Angela Benedetti called for a reinstatement of the Carrot Law in the city. The Council responded by passing other measures including the closing of liquor sales in some neighborhoods at 1 am. In April 2009 Councilwoman Benedetti reported that these measures were not working and called again for reinstatement of the Carrot Law, this time by neighborhoods. She also noted that many establishments were getting around the 3 am closing by reestablishing themselves as liquor stores, grocery stores and corner shops.

Sometime around March 2011 the Mayor established a study committee including his office, the metropolitan police and health services to examine the relationship between the sale and consumption of liquor and violence in different parts of the city and at different hours. After four months, the Mayor issued Decree 263 prohibiting liquor sales by stores in seven designated districts ('localidades'; there are 20 total in Bogota but these seven accounted for 64% of homicides in the period studied) between 11 pm and 10 am for the long holiday weekend of 24 27 June; bars could still remain open until 3 am. On June 28, 2011 the Mayor reported that the Decree was a resounding success, with significant declines in murders, car accidents and even general accidents. The Decree was extended. On 12 July the Mayors Office announced that sanctions for violation of the Decree had increased significantly. The Bogota Chamber of Commerce reported that 1,374 liquor selling establishments had filed to change their status to bars in order to sell liquor after 11 pm. The Mayors Office announced that they would study the gradual extension of Decree 263 to other parts of the City.

On 17 July the Editor In Chief of El Tiempo newspaper published an editorial saying that those who consume alcohol responsibly should still give up their right to purchase alcohol at late hours so that society could benefit from the decrease in irresponsible alcohol consumption. On 24 July the Mayors Office noted that the Decree would be extended to 24 August and modified the Decree, saying that any business that sold alcohol had to close from 9 pm

10 am.

17 August 2011 the Mayors Office reported to the City Council that Decree 263 had contributed to a 26% reduction in homicides, a 48% reduction in traffic deaths , a 19% reduction in accidents, and a 36% reduction in traffic accidents. The Mayor also reported that 6,585 of 80,000 establishments that fell under the Decree had violated the Decree and were either fined or closed.

In March 2012 the Bogota Chamber of Commerce reported in its survey that 45% of respondents said that insecurity in the city was increasing and called on the Mayor to keep the restrictions on alcohol sales.

On 2 August 2013, the National Commerce Federation, the National Association of Businesses of Colombia, the Colombian Association of the Liquor Industry, and the Colombian Association of Importers of Liquor and Wine announced that the liquor restrictions had led to a 25% reduction in sales by small businesses which also sold liquor in these neighbors. These organizations asked the Mayor to rescind the Decree and establish security study groups on which they would have representation.

On August 5 and 6 businessmen protested in the center of the city against the discriminatory nature of the restrictions, probably in hopes that businesses in non-affected parts of the city would join in the protest, but there are no press stories suggesting that the concerns were becoming city-wide.

Trigger

On 8 August the Mayors Office did not show up for a morning meeting with representatives from the seven affected neighborhoods. The Decree was scheduled to be modified or extended by the City Council on the next day.

Follow-On

On 2 September the Mayors Office issued Decree 374 which prohibited the sale of liquor in establishments in 469 neighborhoods of Bogota between 9 pm and 10 am for 13 of the 30 days in September.

Potential Political Entrepreneurs

The following entrepreneurs were found through the analysis: Councilwoman Benedetti; Mayor Gustavo Petro; Juan Ernesto Parra, National Director of the National Federation of Businessmen (Federación Nacional de Comerciantes); Camilo Llinás, Director General of the National Association of Businesses of Colombia, Bogotá section; and Francisco Alvarez Muñoz, President of the Association of Stores of Bogota (Asociación de Tiendas de Bogotá).

New Data Sources

The following additional data sources were identified: NoticiasRCN.com, Caracol.com.co, and ElEspectador.com.

3.3.4 Conclusion

After collecting 2597 comments over 7 months, we believe that the Reverse OSI interface is successful at allowing investigation of the outputs of machine learning algorithms. Given the number of comments per user, we believe that this interface scales reasonably well to support many investigations of many events. Table 3.1 shows the number of comments contributed by the top 5 most prolific commenters throughout the study period. As future work, to support sharing links to commented events, we could implement persistent URLs to

particular events or comments. This could allow for collaboration and conversation among users commenting on the same event or for distribution of URLs as tasks for crowd sourcing.

Table 3.1: Top 5 most prolific users

User ID	Number of Comments
30	926
14	697
8	490
22	225
29	171

Chapter 4

Interactive Model Building

Our first interface, the warning and audit trail visualizations, provided a basic view of the state of EMBERS outputs. The Reverse OSI capability had more to do with analysis of missed predictions and how to improve forecasting performance in the future. Our third interface, described in this chapter, aims to fundamentally improve how machine learning algorithms are constructed/tuned in EMBERS.

4.1 Requirements Analysis

We focus on machine learning models that deal with tweets. We aim to design an interface that enables a user to identify a set of tweets that, taken together, are a precursor for a certain event. Users begin by picking a GSR event and consider the tweets with time-stamps before this event. They mark those tweets that they recognize as precursor tweets to the selected GSR event, implying that there is some signal in those tweets that will help predict their chosen event. Then, server-side, the website will train a model to classify that set of tweets as an event of the same type as the GSR event the user chose (recall that GSR events

have several types: population, violence, event type). This will generate a model but we need to provide feedback to users about how well the new model works and whether it requires further tweaking. These new models will work much like the models in the official EMBERS suite. In other words, they ingest tweets for some discrete window of time, perform some calculations, and then either make predictions, wait for the next window of time, or decide that there is no event to predict. We can use the model from the user's input in the same way. To start with, we can decide if the model the user made would have predicted the original GSR event that was used to make this model. To do this, we take the tweets leading up to the time of that event and evaluate them through the new, user-generated model. Then, we can see if that model generates a prediction that matches the event that they used to generate the model. This is the most basic evaluation check. If the model does not predict the event it was meant to predict in the first place then it may well require further more work. Secondly, since the model works on streams of tweets, we can evaluate it on historical tweets and compare against a longer window of the GSR. Ideally, the model will predict events other than only the event that it was trained on and not be overly susceptible to overfitting. By iterating on the process of marking tweets as precursors, retraining the model , and evaluating the models effectiveness, users with domain specific knowledge can make models that can potentially be added to the EMBERS suite.

4.1.1 Interactive Machine Learning

Much work has been done in involving users in the process of building classifiers. Classifiers pose the question which group (or class) a given item belongs to. Ware et al. [20] developed an interface for users to build classifiers by manually lassoing elements of each class in dimensionally reduced space. They found that not only were user-defined classifiers easier to understand but also they were competitive with machine-built classifiers when classification

could be made by few features. At the same time there is work being done to explain why predictive algorithms come to the conclusions they do. Malik et al. [21] explain geospatial crime rate predictions as a choropleth. And there has been increasing formalization of ways to explain complex analytics algorithms in ways that benefit sense making and performance tuning [22].

There has been activity in providing interfaces that let users who are potentially untrained in machine learning techniques improve the quality of learned models. In many cases this involves users helping to shape input training sets into machine learning algorithms. This kind of work is most closely related to ours which provides interactions to guide predictive machine learning algorithms rather than to visualize the results of them. Mühlbacher and Piringer had users help the machine build regression models by having a say in feature subset selection [23]. Krause et al. [24] developed a system that allows users to interactively build classifiers by choosing between feature selection and classification algorithms.

Figure 4.1: Screenshot of interactive model building interface in EMBERS. Used with permission of Dr. N. Ramakrishnan, 2015.

4.2 Design and Implementation

To give historical context to the user, the top center of the page contains a graph that displays tweet volume over a window of time. For the prototype website shown in Figure 4.1, we made this window span a one month timeframe. For this chart each unit along the x-axis is a thirty minute window of time and each vertical region is the volume of tweets sent in that time, centered to give the waveform look. As shown, usage of Twitter generally peaks around lunch time and then again after dinner before dying down in the middle of the night. This chart gives some context into what time period and how much Twitter data there is to work with. Above the waveform, circles indicate GSR events. These circles are placed along the x-axis for the time that they occurred. To the left there are some controls for filtering the types of events that we are looking at To the right there is room for displaying the information about a GSR.

4.2.1 Building the Tweet Set

Once an event is chosen, the user can work on building a set of precursor tweets for that event. The current set of tweets is displayed in a tabbed pane with each tab representing an addition by a certain rule. Rules are queries into the database of tweets that governs this set. To kickstart this process, our system injects a set of tweets that that have a high probability of relating to the event. These tweets become the first rule for the precursor set as shown in Figure 4.2.1. In most cases an initial set consists of a few hundred tweets which represent a best guess at gathering tweets that are relevant. The user can generate their own rules by adding tweets to or removing tweets from this set until it represents their understanding of what should and should not be a precursor and the model generated from this set performs to their satisfaction.

For each addition or subtraction the user makes from the set, a rule is added to the list of rules. To give the user an idea of the impact of their actions, each rule in the list has a badge indicating the number of tweets affected by that modification. So that no work is lost, existing rules can be removed and readded to the set. Each rule has a close button which will move the rule to a list of deleted rules and apply that rule's inverse. Rules in the list of deleted rules can be reinstated via a button on their list item. Several interactions result in the introduction of a new rule:

- Remove all tweets with a certain word. Selecting any word invokes a context menu which offers the option to remove all tweets containing the selected word.
- Remove a single tweet. Each tweet can be removed from the set via an X button.
- Add more tweets with a certain word. The context menu for each word also contains an option to add more tweets that contain the given word.
- Add tweets from search. The search tab adds tweets with a given phrase to the set.

Any interactions that make additions to the set obtain tweets from the collection of historical tweets for this time period. The time period is initially set from the earliest date of tweets in the initial set of precursor tweets to the date of the event. This range can be changed by editing the timestamps in the special date range rule in the rules list or by dragging the date range selector in the tweet volume graph. When rules make additions to the set, new tweets are added in a new tab in the dataset panel. Tabs are given a descriptive name based on the query that added them.

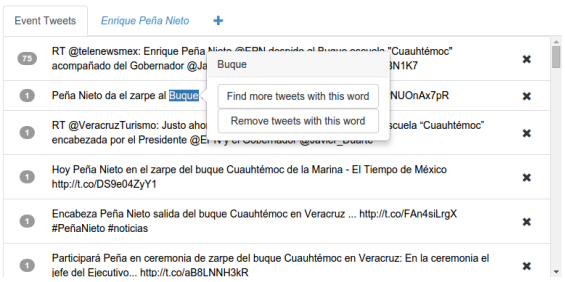


Figure 4.2: The initial tweets appear in the first tab. Selecting any word in a tweet brings up a popover to add new rules based on that word. Tweets added by previous searches appear in their own tab. Used with permission of Dr. N. Ramakrishnan, 2015.

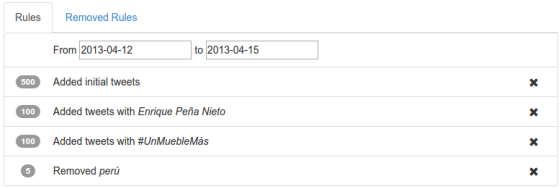


Figure 4.3: The rules list describes the makeup of the tweet set. The first rule is a special date range rule which can be modified in place. Each rule has a badge containing the number of tweets affected and a mark for removing it. Removed rules are moved to the removed rules tab. Used with permission of Dr. N. Ramakrishnan, 2015.

4.2.2 Assessing Predictive Models

Each time a user adds or removes a rule, the interface communicates the set of tweets to the model builder which begins the somewhat time consuming process of generating a model that predict events from tweets. Then historical tweets are fed into the model as a stream and any predictions made by the model are visualized. Predictions are visualized as triangles above the circles that represent ground truth events. Statistics for the model including the precision and recall of the predictions against ground truth events of the same type are shown in the statistics tab. In this way a user can determine:

1. Does the current model predict the event for which the tweets are a precursor?

2. How well does the current model predict other events of this type?

4.2.3 Tweaking the Model

A user has several ways to change the model after assessing its success. Continuing to iterate on the set of rules by adding or removing more tweets will continue to update the view of predictions so the user can get closer and closer to the results they want. Also, a user can add another pair of event and set of precursor tweets to the input to the model builder. When a new event is chosen, it comes with a brand new set of rules empty of everything except the initial set of tweets for the new event and the automatically computed date range rule. The circles for other events with user-built precursor tweet sets are highlighted to indicated that they have rules associated with them. Selecting highlighted circles for events other than the currently selected one will change the view to show the rules and tweet set for that event.

Selecting any prediction triangle will display the tweets used in that prediction. This gives the user some idea of why the prediction was made. In the case of incorrect predictions, the user can generate new rules for each event with rules by removing tweets with a certain word or adding more with a certain word as with the tweets in the input set.

4.3 Evaluation

The following case study provides an example of how the website could be used to leverage the expertise of a user to build a predictive model. This serves as an evaluation of the fitness of the design for iteratively building a model.

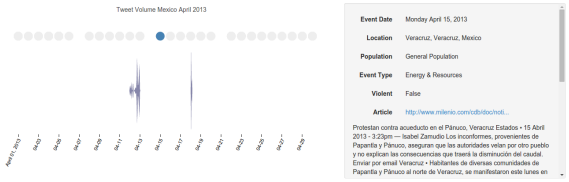


Figure 4.4: Interactive model building interface. Used with permission of Dr. N. Ramakrishnan, 2015.

4.3.1 Case Study

To demonstrate how the interface works, consider the following example. The data used by the prototype site includes Twitter data from Mexico for the month of April 2013. In addition to this data, the GSR events for that month are loaded above the chart of Twitter volume as shown in Figure 4.1. The task for the user is to investigate an event that was missed by EMBERS and train a model to hopefully forecast this type of event in the future. After choosing the event of interest, the Twitter volume chart updates to show the initial set of tweets related to that event and the main details from the GSR event along with the article contents of the GSR event as shown in Figure 4.3.1. The user reads the GSR article and finds out that this event is triggered by lack of transparency in various levels of government that govern the water supply for the communities near Veracruz. To train the model, the user needs to construct a set of tweets that seem to be precursors to this event. She starts reading through the initial set of tweets to get an idea of what tweets the website thinks are relevant to this event. She notices that many tweets mention Mexican president Enrique Peña Nieto but that these tweets deal with a ceremony involved with his visit to a particular ship *buque* at a Naval facility. Since this does not have much to do with the complaints of the residents of Veracruz, she clicks on *buque* to bring up a context menu for that word from which she chooses to remove tweets that contain that word as shown in Figure 4.3.1. This action adds a new rule to the set that indicates that 237 tweets with the word *buque* were removed, as shown in Figure 4.3.1.

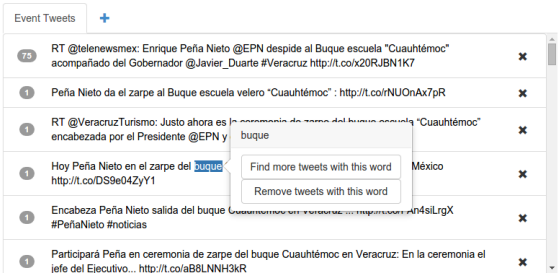


Figure 4.5: Word-specific context menu that can request more tweets or remove all tweets that contain this word. Used with permission of Dr. N. Ramakrishnan, 2015.

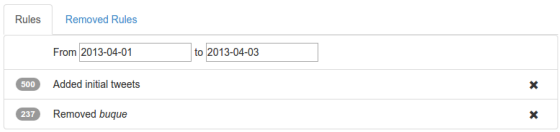


Figure 4.6: Example of rule that removes tweets with badge indicating how many were removed. Used with permission of Dr. N. Ramakrishnan, 2015.

Since the article description explains that citizens are expressing discontent with the country's water agency (*Comisión Nacional del Agua*), she decides to try to add tweets with that phrase. To do this, she selects the plus sign tab as shown in Figure 4.3.1, enters *Comisión Nacional del Agua*, and clicks the search button. This searches for tweets in the date range listed in the rules list for tweets with that phrase. Tweets that contain this phrase are added in their own tabbed pane to the set of tweets that are precursors to this event. In this case there is only one so the user reads it and decides it is fine to stay in the set of precursor tweets. At this point, she is ready to try out if the precursor set of tweets does a good job of predicting events so she clicks the Run button. At this point, the website sends the current set to the server which trains a predictive model on that set. When the training is done, predictions are sent back to the client where the user and inspect the predictions. She can compare these predictions with GSR events of the same type and decide if the newly trained model is performing well. Now, she can continue to iterate by adding or removing tweets to the set of precursors, retraining the model, and evaluating its predictions.

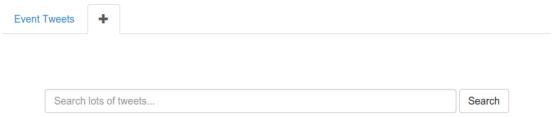


Figure 4.7: The search for more tweets search tab. Used with permission of Dr. N. Ramakrishnan, 2015.

4.3.2 Conclusions and Future Work

This case study suggests that this interface is a reasonable design for the task. It allows the user to act on their intuition as they build the model. Despite this, it would be beneficial to evaluate the tool in a more controlled setting with more diverse users.

One of the biggest problems with this tool is that training the model on a user-defined set of tweets takes a long time. Though there is opportunity for algorithm research on speeding up this process, there are ways to mitigate this problem. Work is being done on providing the user a best guess answer to questions that will take a long time to compute [25]. Together with modifications to the training algorithm, it could be possible to have something to show a user in near real-time while on the backend the algorithm is incrementally solving the problem and updating the view as better answers become available. If there is faster feedback with indications of how correct it is likely to be compared to the complete answer, such a facility could help users decide on the next step while the answer is still emerging and discard wrong paths more quickly.

Another next step is to integrate websites like this into the broader EMBERS system. Other sites that deal with GSR events could link into this site to start building models from events that were being inspected elsewhere. Further, models generated from this site could be used in real-time production of event predictions along with sites to track how well those models

are doing. In this way, we could distribute the task of establishing a number of statistical models across many expert users.

Chapter 5

Conclusion

This thesis described three interfaces for interacting with an open source indicators forecasting system. First, the prediction overview page summarized the state of EMBERS, a distributed suite of statistical models that ingest hundreds of gigabytes of open data to forecast events. This page uses cross linked charts to allow users to gather insights about trends over time, place, or other dimension in the forecasts EMBERS has made. Closely related to this is the audit trail page which condenses large JSON audit trails that encode the provenance of a single prediction into a more easily digestible visualization. Because of the differences in audit trail format across different models that make predictions for EMBERS, this page has a modular format so that the details area can be different for each model. The Reverse OSI website allowed for structured commenting on thousands of GSR events to investigate the characteristics of events that EMBERS failed to predict and to attempt to find new sources that would have helped predict them. This site allowed for more user input into the workings of EMBERS than the prediction and audit trail visualizers since model developers and social scientists used the comments gathered on this site to improve EMBERS functionality. Finally, the prototype for the interactive model building web page

allows users to build statistical models for predicting events with the goal of incorporating these models into EMBERS production environment. This site had the smallest turnaround time from user input to feedback about how useful that input was.

As future work, the Reverse OSI and interactive model building sites could be opened up for crowdsourcing through a platform like Mechanical Turk. This would give us the opportunity to bring in greater domain input into EMBERS. It would require some modifications to ensure that users without knowledge of the terminology and practices of EMBERS and the OSI project can understand what is needed by these sites. Further future work could investigate expanding the domains of EMBERS predictions. EMBERS currently focuses on domains that are interesting to policy makers and intelligence analysts. With functionality like the interactive model building website, users could make predictions for domains relevant to their needs such as the price of consumer goods or flights.

Bibliography

[1] S. Asur, B. Huberman, *et al.*, "Predicting the Future with Social Media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 1, pp. 492–499, IEEE, 2010.

[2] A. Culotta, "Towards Detecting Influenza Epidemics by Analyzing Twitter Messages," in *Proceedings of the First Workshop on Social Media Analytics*, pp. 115–122, ACM, 2010.

[3] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[4] 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*, pp. 1799–1808, ACM, 2014.

[5] A. Doyle, G. Katz, K. Summers, C. Ackermann, I. Zavorin, Z. Lim, S. Muthiah, P. Butler, N. Self, L. Zhao, *et al.*, "Forecasting Significant Societal Events Using The Embers Streaming Predictive Analytics System," *Big Data*, vol. 2, no. 4, pp. 185–195, 2014.

[6] A. Doyle, G. Katz, K. Summers, C. Ackermann, I. Zavorin, Z. Lim, S. Muthiah, L. Zhao, C.-T. Lu, P. Butler, *et al.*, "The EMBERS Architecture For Streaming Predictive Analytics," in *Big Data (Big Data), 2014 IEEE International Conference on*, pp. 11–13, IEEE, 2014.

[7] H. Llorens, L. Derczynski, R. J. Gaizauskas, and E. Saquete, "TIMEN: An Open Temporal Expression Normalisation Resource.," in *LREC*, pp. 3044–3051, 2012.

[8] M. M. Bradley and P. J. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings," tech. rep., Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, 1999.

[9] S. H. Bach, M. Broecheler, B. Huang, and L. Getoor, "Hinge-Loss Markov Random Fields and Probabilistic Soft Logic," vol. arXiv:1505.04406 [cs.LG], 2015.

[10] L. Zhao, F. Chen, J. Dai, T. Hua, C.-T. Lu, and N. Ramakrishnan, "Unsupervised Spatial Event Detection in Targeted Domains with Applications to Civil Unrest Modeling," *PLoS ONE*, vol. 9, p. e110206, 10 2014.

[11] M. Bostock, V. Ogievetsky, and J. Heer, "D3: Data-Driven Documents," *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011.

[12] T. Munzner, *Visualization Analysis and Design*. CRC Press, 2014.

[13] S. Muthiah, B. Huang, J. Arredondo, D. Mares, L. Getoor, G. Katz, and N. Ramakrishnan, "Planned Protest Modeling in News and Social Media," *Innovative Applications of Artificial Intelligence*, 2015.

[14] T. Hua, C.-T. Lu, N. Ramakrishnan, F. Chen, J. Arredondo, D. Mares, and K. Summers, "Analyzing Civil Unrest through Social Media," *Computer*, vol. 46, pp. 80–84, Dec 2013.

[15] J. Weng and B.-S. Lee, "Event Detection in Twitter," *ICWSM*, vol. 11, pp. 401–408, 2011.

[16] F. Chen, J. Arredondo, R. P. Khandpur, C.-T. Lu, D. Mares, D. Gupta, and N. Ramakrishnan, "Spatial Surrogates To Forecast Social Mobilization And Civil Unrests," in *Position Paper in CCC Workshop on From GPS and Virtual Globes to Spatial Computing-2012*, 2012.

[17] P. Chakraborty, P. Khadivi, B. Lewis, A. Mahendiran, J. Chen, P. Butler, E. O. Nsoesie, S. R. Mekaru, J. S. Brownstein, M. Marathe, *et al.*, "Forecasting a Moving Target: Ensemble Models for ILI Case Count Predictions," in *Proceedings of the 2014 SIAM International Conference on Data Mining. Proceedings. Society for Industrial and Applied Mathematics*, pp. 262–270, 2014.

[18] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[19] K. A. Cook and J. J. Thomas, "Illuminating The Path: The Research And Development Agenda For Visual Analytics," tech. rep., Pacific Northwest National Laboratory (PNNL), Richland, WA (US), 2005.

[20] M. Ware, E. Frank, G. Holmes, M. Hall, and I. H. Witten, "Interactive Machine Learning: Letting Users Build Classifiers," *International Journal of Human-Computer Studies*, vol. 55, no. 3, pp. 281–292, 2001.

[21] A. Malik, R. Maciejewski, S. Towers, S. McCullough, and D. Ebert, "Proactive Spatiotemporal Resource Allocation and Predictive Visual Analytics for Community Policing and Law Enforcement," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 20, pp. 1863–1872, Dec 2014.

[22] T. Muhlbacher, H. Piringer, S. Gratzl, M. Sedlmair, and M. Streit, "Opening the Black Box: Strategies for Increased User Involvement in Existing Algorithm Implementations," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 20, pp. 1643–1652, Dec 2014.

[23] T. Muhlbacher and H. Piringer, "A Partition-Based Framework for Building and Validating Regression Models," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 12, pp. 1962–1971, 2013.

[24] J. Krause, A. Perer, and E. Bertini, "INFUSE: Interactive Feature Selection for Predictive Modeling Of High Dimensional Data," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 20, no. 12, pp. 1614–1623, 2014.

[25] D. Fisher, I. Popov, S. M. Drucker, and mc schraefel, "Trust Me, I'm Partially Right: Incremental Visualization Lets Analysts Explore Large Datasets Faster," in *Proceedings of the 2012 Conference on Human Factors in Computing Systems (CHI 2012)*, ACM Conference on Human Factors in Computing Systems, May 2012.