ORIGINAL RESEARCH

FORECASTING SIGNIFICANT SOCIETAL EVENTS USING THE EMBERS STREAMING PREDICTIVE ANALYTICS SYSTEM

Andy Doyle,[1] Graham Katz,[1] Kristen Summers,[1] Chris Ackermann,[1] Ilya Zavorin,[1] Zunsik Lim,[1] Sathappan Muthiah,[2] Patrick Butler,[2] Nathan Self,[2] Liang Zhao,[2] Chang-Tien Lu,[2] Rupinder Paul Khandpur,[2] Youssef Fayed,[3] and Naren Ramakrishnan[2]

Abstract

Developed under the Intelligence Advanced Research Project Activity Open Source Indicators program, Early Model Based Event Recognition using Surrogates (EMBERS) is a large-scale big data analytics system for forecasting significant societal events, such as civil unrest events on the basis of continuous, automated analysis of large volumes of publicly available data. It has been operational since November 2012 and delivers approximately 50 predictions each day for countries of Latin America. EMBERS is built on a streaming, scalable, loosely coupled, shared-nothing architecture using ZeroMQ as its messaging backbone and JSON as its wire data format. It is deployed on Amazon Web Services using an entirely automated deployment process. We describe the architecture of the system, some of the design tradeoffs encountered during development, and specifics of the machine learning models underlying EMBERS. We also present a detailed prospective evaluation of EMBERS in forecasting significant societal events in the past 2 years.

Introduction

Anticipatory intelligence is considered to be one of the next frontiers of "big data" research, wherein myriad data streams are fused together to generate predictions of critical societal events. One of the promising themes in this space is the idea of harnessing open-source datasets to identify threats and support decision making for national security, law enforcement, and intelligence missions. Early Model Based Event Recognition using Surrogates (EMBERS)[1] is an anticipatory intelligence system for forecasting socially significant population-level events, such as civil unrest incidents, disease outbreaks, and election outcomes, on the basis of publicly available data. EMBERS is supported by the Intelligence Advanced Research Project Activity (IARPA) Open Source Indicators (OSI) program.

The classes of events EMBERS is designed to forecast include influenza-like illness case counts, rare disease outbreaks, elections, domestic political crises, and civil unrest (we focus in this article primarily on civil unrest). For civil unrest, EMBERS produces detailed forecasts about future events, including the date, location (to within a city resolution), type of event (e.g., whether it is a protest for wages or a protest for safety), and protesting population (e.g., educators, factory workers, doctors), along with uncertainties involved in the forecasts. It has been operational and delivering predictions (and continues to) since November 2012. The system processes a range of data, from high-volume, high-velocity, noisy open-source media such as Twitter to lower-volume, higher-quality sources, such as economic indicators. Much of the system is designed to look for precursor signals in social media streams and use these

[1] *CACI Inc., Lanham, Maryland.*
[2] *Virginia Tech, Arlington, Virginia.*
[3] *BASIS Technology, Herndon, Virginia.*

indicators to drive statistical and machine learning algorithms that generate the predictions.

Three key considerations motivated the design of EMBERS. First, the EMBERS system architecture was designed to support collaboration from the outset. The team composition involves eight research universities and two industry partners contributing diverse expertise in computer science, machine learning, disease modeling, social science, linguistic processing, and systems integration. This diverse team required a highly distributed and loosely coupled functional architecture, allowing team members to develop components for the system without worrying about dependencies among them. Additionally, the vast majority of the processing is performed on continuous streams of data that need to be processed in near-real-time. To address these needs, the system is composed of many simple independent components strung together in a pipes-and-filters architecture.[2] In particular, EMBERS is built on a simple message-passing design that requires few shared dependencies between components and no shared infrastructure.

Second, rather than develop one monolithic prediction model, EMBERS takes the approach of developing multiple machine learning models. Each of these models is tuned for high precision so that the union of their outputs achieves (high) recall objectives. A key fusion/suppression stage in EMBERS enables us to combine the selective superiorities of the underlying models.

Third, EMBERS is designed to capture the systematic transduction of raw data into final warnings (see Fig. 1). A data taxonomy in EMBERS supports the transformation of raw data, into enriched data, into surrogate information, and finally to the warnings or alerts. Following this data taxonomy in reverse, we have the ability to reconstruct an "audit trail" of every warning issued by EMBERS. Furthermore, EMBERS has the facility to conduct ablation studies wherein specific data sources can be eliminated to understand their impact on the final predictions.

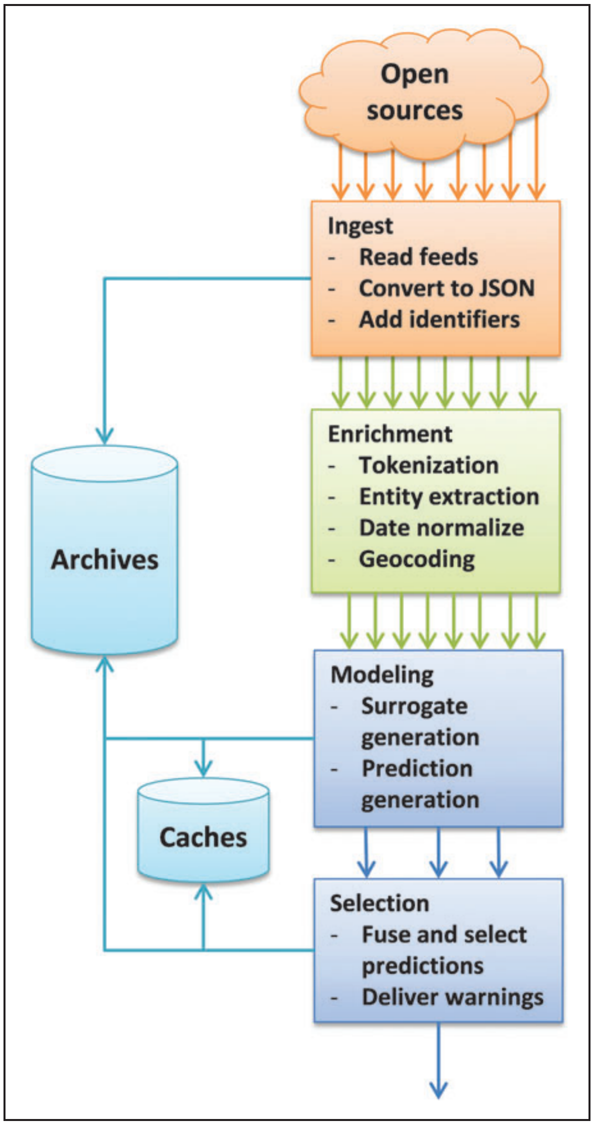


FIG. 1. Early Model Based Event Recognition using Surrogates (EMBERS) system components.

The focus of EMBERS is on countries of Latin America (with recent expansion to the Middle East and North Africa [MENA] region). Predictions made by EMBERS are evaluated monthly against ground truth data (called the gold standard report [GSR]) created by human analysts at MITRE. Currently in the third year, the system has met many of its target objectives for the first 2 years. In addition, the system has encountered a number of real-world events that were unexpected and significant in size. EMBERS has been able to find indicators of these events in social media content and produce predictions that matched both the timing of the events and their trajectory in terms of size and intensity. Two key examples are the series of protests in Brazil in June 2013 and the violent student-led protests in Venezuela in February 2014.

"TO ADDRESS THESE NEEDS, THE SYSTEM IS COMPOSED OF MANY SIMPLE INDEPENDENT COMPONENTS STRUNG TOGETHER IN A PIPES-AND-FILTERS ARCHITECTURE."

The key contributions of this article are as follows:

1. We outline the design architecture of EMBERS, paying particular attention to design decisions, tradeoffs, and implementation of EMBERS in a new environment.
2. We describe the multiple levels of data transduction that happen in EMBERS enabling the real-time analysis of massive data streams.

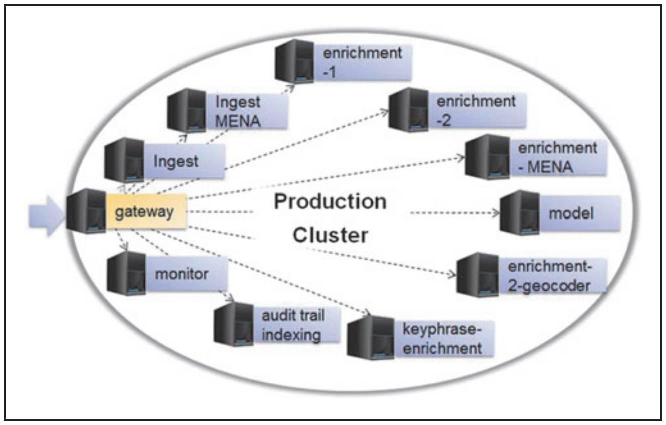


FIG. 2. Layout of the EMBERS cluster.

3. We identify the numerous predictive models in EMBERS and present a detailed performance evaluation of these models.
4. We provide a retrospective commentary on events that were forecast correctly by EMBERS and how EMBERS can be utilized by policy makers.

Related Work

Significant work exists in the space of streaming big data processing systems and we were motivated to leverage lessons learned from earlier efforts. We survey the most closely related work here. A solution using NoSQL databases, key-value stores (Redis), and high-throughput queues was used in Ref.[3] to process Twitter and bit.ly streams. Distributed streaming algorithms for analyzing people mobility data are presented in Ref.[4] The Resa system[5] supports real-time streaming analytics in the cloud paying particular attention to dynamic additions and removals of data processing operators. This framework has been used for Twitter analytics to detect both frequent and outlier patterns. Santos et al.[6] describe the DiAI system for distributed analytics over the cloud using a Java-based event processing framework. EMBERS is distinguished from these efforts by the diversity and heterogeneity of data sources we seek to harness, the need to support distributed model development without disturbing existing data flows, and by the nature of real-time forecasts it generates as output.

> "THIS PERMITS THE EASY COMPOSITION OF NEW TOPOLOGIES FOR ROUTING DATA AND THE STRAIGHTFORWARD INSERTION OF NEW COMPONENTS INTO THE RUNNING SYSTEM."

Architecture

The EMBERS system consists of four major processing components: Data Ingest, Message Enrichment, Analytic Modeling, and Prediction Fusion. These components are illustrated in Figure 1. While the primary mode of data processing is the analysis of streaming data, the architecture is flexible enough to support batch processing and database-based storage for models that need to aggregate data or support evaluation of past data as part of their model. Parts of the system also perform streaming aggregation of data either in the form of aggregating statistics like keyword counts, or for creating sliding windows of data for time series processing.

The EMBERS system is composed of a set of independent cooperating processes that process multiple data feeds in parallel, eventually feeding models that generate predictions. EMBERS follows the UNIX philosophy,[7] having individual components perform single actions on input data elements (e.g., tokenizing text, normalizing dates, or geocoding messages), adding output of this action to the data structure and publishing it for other consumers; producers and consumers need to know only the endpoint for the data and the field structure of the datum. This permits the easy composition of new topologies for routing data and the straightforward insertion of new components into the running system. It also allows for parallelism: in many cases the same streams of data are used by different models to make predictions. This parallel execution is encouraged by loose coupling of the components and simple broadcast of all the data.

Data transport

Because the EMBERS architecture supports development activity by a diverse team of researchers using a diverse set of implementation strategies, minimizing dependencies was a high priority. Lightweight message passing and open data formats were critical to allowing researchers to experiment and prototype with a variety of tools and data sources and then to seamlessly integrate their work into the running system. Because the publish and subscribe connections between components are configuration driven, new processing steps or parallel processing paths can be added to the system without disrupting existing data flows.

Data is exchanged using JSON messages transmitted over ZeroMQ sockets. Having a small footprint communications library has allowed developers to work independently and only worry about dependencies that had direct relevance to their tasks.

Operational infrastructure

EMBERS is deployed on the commercial Amazon Web Services cloud infrastructure. The current cluster configuration consists of 12 machines with a total of 21 virtual CPUs and 75 GB of RAM. Processes are distributed in the cluster (Fig. 2) according to their resource needs. The cluster is a collection of nodes (EC2 virtual machines) that host a collection of services that read from or write to streams (ZeroMQ queues or S3 files). Cluster setup and deployment are automated: A single command is used to deploy a new cluster or update a build on an existing cluster.

Data provenance and auditing

Each EMBERS message is tagged with a globally unique identifier and a timestamp. Derived messages are tagged with the identifier of the message they are derived from. This chain of messages can be used to trace the processing of a message through the system. Therefore, when a warning is produced by the system, it is already connected to all of the data that produced it. This chain can be accessed in reverse so that the source of each warning can be discovered.

The messages are indexed in Dynamo DB so that the derivation chains can be quickly accessed. The message identifier and its derived identifiers are kept in the index with an offset to the full message in one of the archived data files. This allows for efficient retrieval of messages based on their derived identifiers while minimizing the use of centralized database storage.

This information about data provenance is used to visualize an audit trail for each warning as a way of analyzing the reason why each warning was produced. This application uses the indexes generated from the source data and retrieves key parts of the source messages, such as tweet text or news article, to show the derivation of a warning. For models that aggregate data, such as ones that look for keyword clusters or thresholds, the intermediate data can also be shown. Figure 3 shows the detailed view of a single alert in the EMBERS visualization dashboard. EMBERS also supports an ablation visualizer (Fig. 4) wherein an analyst can selectively remove data sources to determine their impact on a particular alert. This is especially useful when the analyst is interested in assessing the relative contributions of data sources or has reason to impute differing fidelities to them.

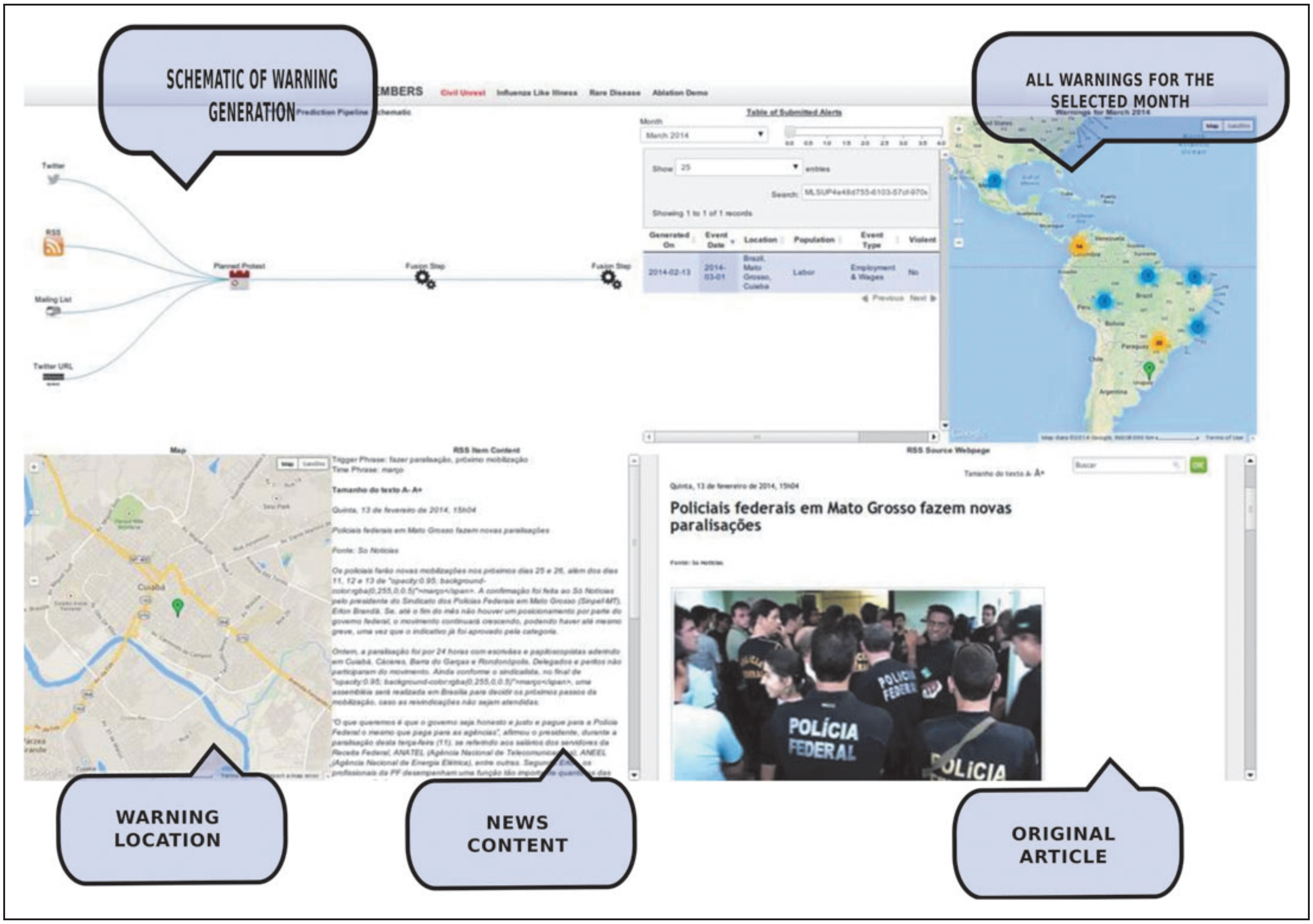


Table of Submitted Alerts

Month: March 2014

Show 25 entries

Search: MLSUP4e48d755-6103-57cf-970x

Showing 1 to 1 of 1 records

Generated On	Event Date	Location	Population	Event Type	Violent
2014-02-13	2014-03-01	Brazil, Mato Grosso, Cuiaba	Labor	Employment & Wages	No

Previous Next

RSS Item Content

Trigger Phrase: fazer paralisação, próximo mobilização
Time Phrase: março

Tamanho do texto A- A+

Quinta, 13 de fevereiro de 2014, 15h04

Policiais federais em Mato Grosso fazem novas paralisações

Fonte: So Noticias

Os policiais farão novas mobilizações nos próximos dias 25 e 26, além dos dias 11, 12 e 13 de "opacity:0.95; background-color:rgba(0,255,0,0.5)">março</span>. A confirmação foi feita ao Só Notícias pelo presidente do Sindicato dos Policias Federais em Mato Grosso (Sinpef-MT), Erlon Brandã. Se, até o fim do mês não houver um posicionamento por parte do governo federal, o movimento continuará crescendo, podendo haver até mesmo greve, uma vez que o indicativo já foi aprovado pela categoria.

Ontem, a paralisação foi por 24 horas com escrivães e papiloscopistas aderindo em Cuiabá, Cáceres, Barra do Garças e Rondonópolis. Delegados e peritos não participaram do movimento. Ainda conforme o sindicalista, no final de "opacity:0.95; background-color:rgba(0,255,0,0.5)">março</span>, uma assembléia será realizada em Brasília para decidir os próximos passos da mobilização, caso as reivindicações não sejam atendidas.

"O que queremos é que o governo seja honesto e justo e pague para a Polícia Federal o mesmo que paga para as agências", afirmou o presidente, durante a paralisação desta terça-feira (11), se referindo aos salários dos servidores da Receita Federal, ANATEL (Agência Nacional de Telecomunicações), ANEEL (Agência Nacional de Energia Elétrica), entre outras. Segundo Erlon, os profissionais da PF desempenham uma função tão importante quanto as das

RSS Source Webpage

Tamanho do texto A- A+

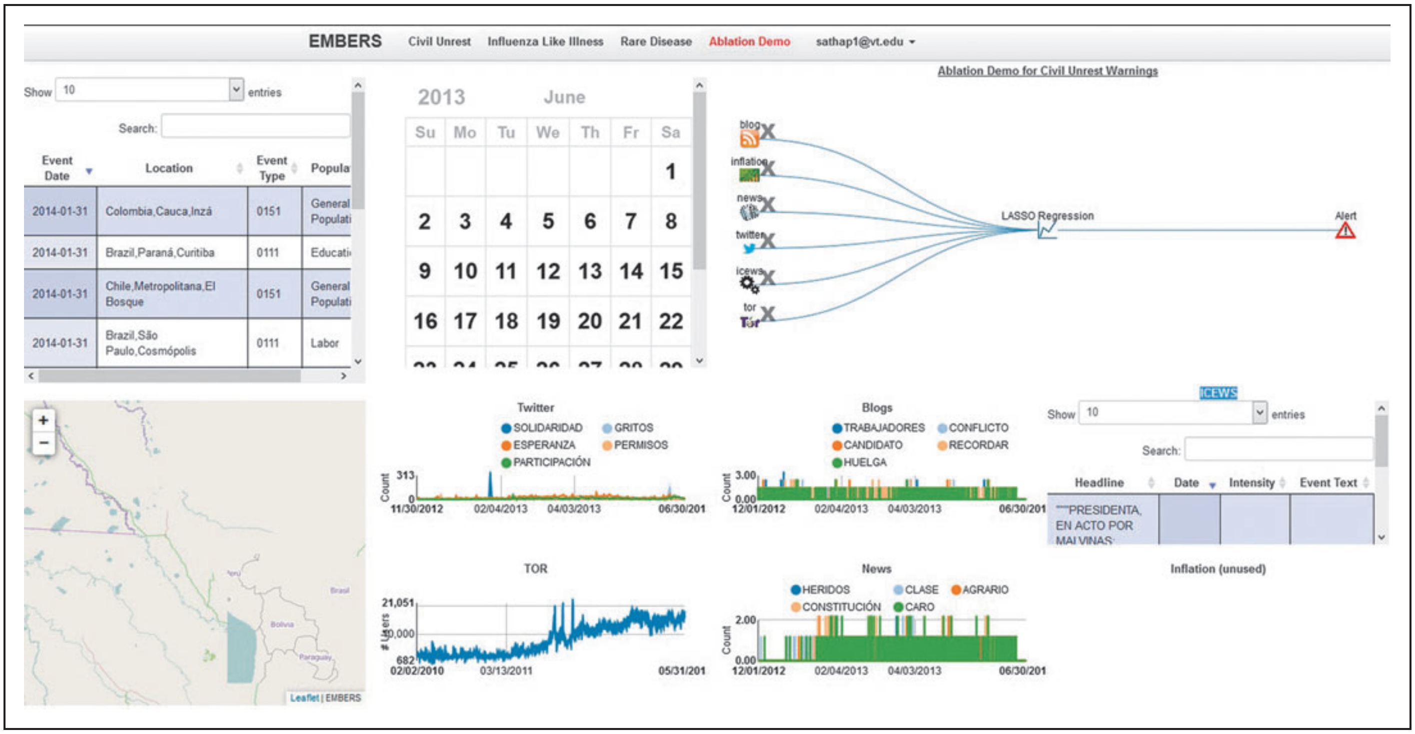
Buscar OK

Quinta, 13 de fevereiro de 2014, 15h04

Policiais federais em Mato Grosso fazem novas paralisações

Fonte: Só Notícias

FIG. 3. An example of the EMBERS audit trail interface.



EMBERS Civil Unrest Influenza Like Illness Rare Disease Ablation Demo sathap1@vt.edu

Show 10 entries

Search:

Event Date	Location	Event Type	Popula
2014-01-31	Colombia,Cauca,Inzá	0151	General Populati
2014-01-31	Brazil,Paraná,Curitiba	0111	Educati
2014-01-31	Chile,Metropolitana,El Bosque	0151	General Populati
2014-01-31	Brazil,São Paulo,Cosmópolis	0111	Labor

ICEWS

Show 10 entries

Search:

Headline	Date	Intensity	Event Text
""PRESIDENTA, EN ACTO POR MALVINAS:			

FIG. 4. EMBERS Ablation visualizer. The top-right section shows the different data sources that are used. Each one of the sources can be selectively removed and its effect on the alerts generated is reflected in the surrounding sections.

Processing Pipeline

In addition to the primary mode of streaming analysis in EMBERS, the architecture also supports batch processing and database-based storage for data aggregation and persistence. This is made heavy use of by the parts of the system that perform streaming aggregation of data (e.g., computing daily keyword counts in real time).

Data ingest

There are approximately a dozen data sources ingested into the EMBERS system, ranging from weekly government reports to Twitter feeds. The largest volume of data is two Twitter feeds that average around 20 messages a second. Each feed is based on queries that are designed to target it to topics and geographic regions that are relevant to the project. Approximately 14,000 Really Simple Syndication (RSS) feeds account for a large amount of text but many fewer messages. Other data sources include curated data such as HealthMap[8,9] alerts and Google Flu Trends data. Most data sources are text based, such as Twitter and RSS feeds, but a significant number, such as Google Flu Trends, are numeric and some are more complex sources such as Global Data Assimilation System (GDAS),[10] which provides climate information derived from satellite data.

The current focus area for EMBERS is Latin America, which means that our data samples are limited to that region. This also means that most of the language processing EMBERS does is for non-English sources, so multilingual processing and analysis is a key aspect of the system. The system currently focuses on processing Spanish, Portuguese, and English sources.

Enrichment

Most of the ingested data passes through a series of enrichment processes to expose information to the downstream models. The highest volume and fastest velocity data sources such as Twitter and RSS feeds are enriched in a pipeline that uses linguistic information to both expand on the content of the textual content of the message and extract information contained in the text into a structured format.

The first enrichment process is basic linguistic processing. This includes tokenization, part of speech tagging, and lemmatization of the individual terms. This step also includes named entity extraction to locate all people, places, organizations, and other expressions such as numbers, dates, and hashtags in the text. The information is added to the original source information and published for the next stage. Basis Technologies RLP and REX products perform these tasks. The core linguistic processing is currently supported for English, Spanish, and Portuguese.

This basic language preprocessing serves as input to subsequent deeper semantic analysis: date normalization, geocoding, and sentiment analysis. Date processing in text is particularly crucial. We based our system for processing English, Spanish, and Portuguese dates on the TIMEN package.[11] This system makes use of metadata, such as the day of publication, and other information about the linguistic context of the date expression to determine for each date expression, what day (or week, month, or year) it refers to. For example, in a tweet produced on June 10, 2014, the occurrence of the term *Friday* used in a future-tense sentence *We'll get together on Friday* will be interpreted as July 13, 2014. Each expression identified as a date by the RLP preprocessor is normalized in this way.

Documents are geocoded with a specification of the geographical focus of the text—specified as a city, state, country triple. We make use of different geocoding methodologies for geocoding news/blogs and for geocoding Twitter postings. For tweets, a trained classifier is used to include features from the tweet, such as the user's profile location and the place as tagged by Twitter along with the tweet text. For tweets that include exact locations, typically from a smartphone's location sensors, the given location is used. It is worth noting that only a small percentage of tweets are tagged with an exact location. In a recent study,[12] just over 2% of all tweets were geotagged, with slightly more than 1.5% having an exact location. For longer form textual sources, typically news articles or blog posts, a more complex system is used to resolve the most likely place name from multiple location entities extracted from the text. For this purpose, a rule-based system using probabilistic soft logic[13] is used.

"THIS BASIC LANGUAGE PREPROCESSING SERVES AS INPUT TO SUBSEQUENT DEEPER SEMANTIC ANALYSIS: DATE NORMALIZATION, GEOCODING, AND SENTIMENT ANALYSIS."

The final stage in the enrichment process is sentiment analysis. The ANEW[14] lexicon is used to derive a three-dimensional sentiment score based on the ANEW scores for the matching tokens. The matching terms and the aggregate scores are provided. An existing Spanish translation of the lexicon[15] was adopted for Spanish text and the lexicon was translated into Portuguese by the EMBERS team.

Part of the enrichment process also identifies information such as URLs referenced in tweets that then become new ingest sources. These sources are fed back into a separate content collection process that gathers this derived data (and subsequently sends it through the enrichment pipeline). These derived sources have become increasingly important since they provide more in-depth content than is supported by social media activity. Experiments have been run looking at similar sources such as photos or videos posted in tweets or to similar social media sources such as Instagram.

Analytic modeling

The process of generating predictions about future events comes from a set of models that operate independently of each other; these are described in detail in the section Forecasting Algorithms. Each model consumes one or more enriched data sources, with some models focusing on Twitter and others leveraging any enriched text-based data source, and still others working on combinations of data sources. This open-ended approach to modeling means that the same data is often evaluated by multiple models and that models can easily expand the data sources and features they consider. Over time, models have been added and removed or adapted to new data sources. Figure 5 illustrates the progressive data reduction that happens across the processing stages in EMBERS. It depicts the volume of data ingested in our RSS feed over the month of April (left scale), and alongside this the number of messages selected picked out as having promising keywords (right scale) as well as the number of warnings (alerts) produced by the models (right).

Data and processing volume

Currently, the system ingests about 19.2 GB of data per day. There are two large-volume Twitter feeds that produce approximately 40–60 and 20–40 messages per second, respectively. The Twitter feeds especially have highly variable volume. Recently, there have been peaks of around 200 messages per second from the existing feeds based on activity related to the World Cup Soccer matches. About 4.6 million messages a day in total are ingested into the system.

When the raw feeds are enriched, the size of the data is about 40 GB per day, which is analogous to the expansion often seen when data is added to a full-text index. Of these 4.6 million messages, about 350 a day are selected as significant by one of the models. From those messages, around 50 warnings a day are generated.

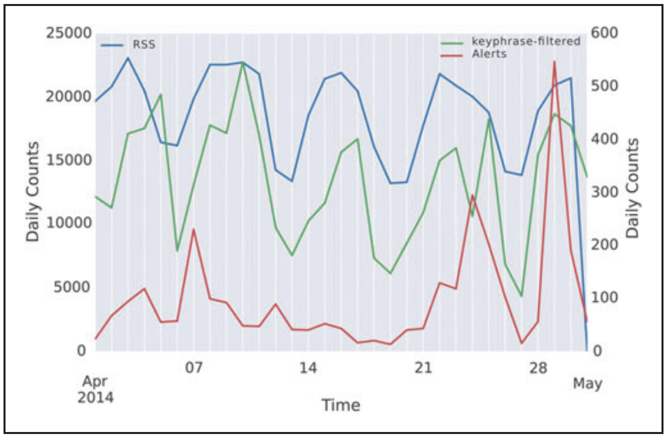


FIG. 5. Data reduction from Really Simple Syndication (RSS) feeds to alerts in an EMBERS civil unrest forecasting model.

The current data mix includes sources from Latin America and the MENA region. As can be seen in Figure 6, Twitter sources account for the vast majority of the data in terms of individual messages. However, looking at the input from the standpoint of data volume, the more text-intensive resources such as RSS feeds and data derived from Tweet URLs provide a much larger share of the data. Figure 6 shows the current breakdown of data by number of messages and by data size.

Resource utilization

While the EMBERS system processes a relatively small sample of overall Twitter traffic and other available data sources, we can estimate the processing efficiency of the current configuration and extrapolate this to estimate the resources needed to process a larger volume of data. The current processing cluster consists of 12 EC2 instances with a total of 21 virtual CPUs and 75 GB of memory. Given that the system processes 4.6 million messages a day and produces roughly 40 GB of data each day, to process 1 million messages it requires about 4.6 CPUs and approximately 16 GB of memory given the current configuration. This includes all EC2 resources allocated to the system, but current utilization of those resources varies, and so this should be considered a conservative estimate. Current CPU utilization is between 20% and 40%, and memory utilization is typically around 75%. Using this as a low-end estimate yields 1.4 CPUs per million messages and 12 GB of memory.

> "USING THIS AS A LOW-END ESTIMATE YIELDS 1.4 CPUS PER MILLION MESSAGES AND 12 GB OF MEMORY."

Forecasting Algorithms

We now outline some specific models used in EMBERS, paying specific attention to their underlying assumptions, data sources, and scenarios of applicability.

Planned protest

Many civil unrest events are planned and organized through calls-for-action by opinion and community leaders who galvanize support for their case. The planned protest model aims at detecting such civil unrest events from traditional media (e.g., news pages, mailing lists, blogs) and from social media (e.g., Twitter, Facebook). The model filters the input streams by matching to a custom multilingual lexicon of expressions such as *preparación huelga*, *llamó a acudir a dicha movilización*, or *plan to strike*, which are likely to indicate a planned unrest event. The phrase matching is done in flexible manner making use of the lemmatized, tokenized output of the BASIS enrichment module, to allow for variation and approximations in the matching. Messages that match are then screened for the mention of a future time/date occurring in the same sentence as the phrase. The event type and population are forecast using a multinomial naive Bayes classifier. Location information is determined using the enrichment geocoders. The phrase dictionary is thus a crucial aspect of the planned protest model and was populated in a semiautomatic manner using both expert knowledge and a simple bootstrapping methodology.

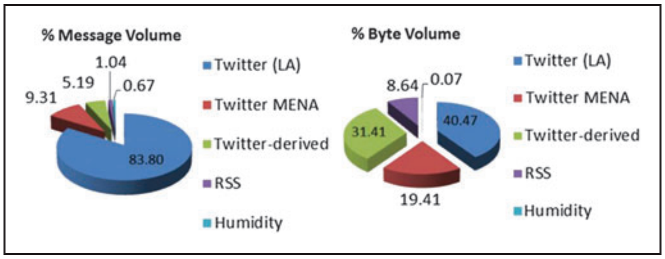


FIG. 6. Data ingest volume in EMBERS.

Initially, a few seed phrases were obtained manually with the help of subject matter experts. These phrases were parsed using a dependency parser and the grammatical relationship between the core subject word—*protest*, *manifestación*, *Huelga*, etc.—and any accompanying word was extracted. To extend the initial set of phrases, a set of sentences containing a subject word and a future time/date expression was collected and parsed. This set of sentences was used to expand the set of planned protest phrases by extracting all keyword combinations that have the same grammatical relation with respect to the core subject word. The final set of planned protest phrases is then obtained after a manual revision of the phrases obtained in the last step.

The planned protest model reads three kinds of input messages: standard natural language text (RSS news and blog feeds, as well as the content of web pages mentioned in tweets), microblogging text (Twitter), and Facebook events pages. The RSS feeds and web pages are processed as discussed above. For tweets, in addition to the above processing, we require that the tweet under consideration be retweeted a minimum number of times, to avoid erroneous alerts. (This value is set to 20 in our system.) For Facebook, we use their public Application Programming Interface (API) to search for event pages containing the word protest or its synonyms. Most such Facebook event pages already provide significant information such as the planned date of protest, location (sometimes with resolution up to street level), and population/category of people involved.

Dynamic query expansion

The dynamic query expansion (DQE) model is based on the idea that the causes for protests can be quite varied, and aims to detect emerging conditions for protests by dynamically growing vocabularies of interest. This model relies exclusively on tweets. Given a short seed query, DQE first adopts an iterative keyword expansion strategy to dynamically generate

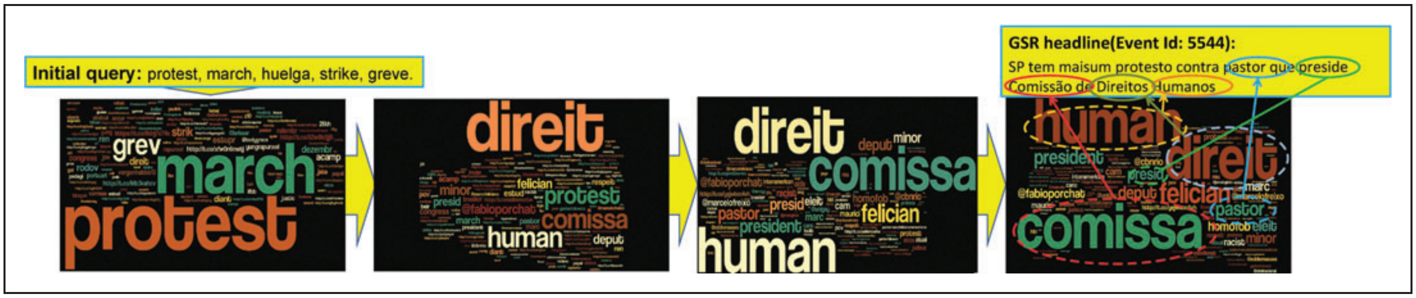

Initial query: protest, march, huelga, strike, greve.
GSR headline(Event Id: 5544):
SP tem maisum protesto contra pastor que preside Comissão de Direitos Humanos
protest
march
grev
direit
human
comissa
felician
pastor
presid
president
deput
minor

FIG. 7. Steps to building a vocabulary using the dynamic query expansion (DQE) model. Beginning from a few general phrases about protests, DQE hones in on keywords relevant to the particular situation being analyzed. As shown, the expanded keywords are pertinent matches to a specific gold standard report event.

a set of extended keywords and tweets pertinent to such keywords. In particular, the seed query consists of a small set of civil unrest-related keywords like “protest” and “march.” In the initial iteration, we extract the tweets matching the seed query, and rank the terms in them by their document frequency-inverse document frequency (DFIDF) weights. Higher-ranked terms are used to trigger the second iteration, continuing the process. The iterations are terminated once the set of keywords and their weights become stable (we have observed that DQE converges in approximately 3–5 iterations; see Fig. 7). The resulting tweets are clustered using local modularity and spatial scan statistics, and tweets in the discovered clusters are used by a classification engine to trigger an alert and to determine the event type and population.

Baseline model

We also developed a maximum likelihood estimate baseline model, making heavy use of a historical database of protests. The idea behind this model is that, even in absence of any explicit signal, the distribution of events that have appeared in the recent past is a good guide to those civil unrest events that will take place in the future. The baseline model makes predictions on the basis of the distribution of “event schema”—frequency in the most recent part of the GSR. An event schema is a combination of a location, an event type, a population, and a day of the week. Some high-frequency schemas can appear as many as 10 times in a 3-month window, but the vast majority of event schemas appear only once. In a typical 3-month interval, two-thirds appear once with the remaining third split evenly between those that appear twice, and those that appear three or more times. Warnings are generated with a minimum threshold of 2, and a 3-month training interval, and issued with a lead time of 2 weeks.

Experimental Results

We outline an exhaustive evaluation of EMBERS alongside multiple aspects. To validate the predictions made by EMBERS, a GSR of protests is organized by a third party (MITRE). The GSR is compiled by human analysts surveying newspapers of record in our countries of interest in Latin America. Given a stream of alerts issued by EMBERS and a stream of events encoded in the GSR, the evaluation problem reduces to bipartite matching between the alerts and events, taking care to ensure that an alert does not get matched to an event in the past. The quality measure for evaluating an (alert, event) matching is defined by a four-component score assessing whether the location, date, population, and event type match between the two records. Each of these components is assessed on a scale of [0, 1] so that a perfect score of 4.0 indicates an exact forecast. The bipartite matching also enables us to define precision and recall in terms of the unmatched alerts and events, respectively.

Table 1 demonstrates that we achieve a quality score of 3.11 with an average lead time of 8.8 days and precision and recall of 0.69 and 0.82, respectively. All but the quality score measure meet or exceed targets set by the IARPA OSI program (OSI’s target for the quality score is 3.25). It is important to

TABLE 1. EMBERS METRICS ACROSS MULTIPLE COUNTRIES

Metric	*AR*	*BR*	*CL*	*CO*	*EC*	*MX*	*PY*	*SV*	*UY*	*VE*	*All*
Quality score	3.2	3.39	2.85	2.86	2.59	3.0	3.27	2.85	3.05	3.01	3.11
Recall	1.0	1.0	0.82	0.59	1.0	1.0	0.65	1.0	1.0	0.84	0.82
Precision	0.55	0.45	0.89	0.94	0.77	0.71	1.0	0.69	0.46	0.73	0.69
Lead time (days)	10.44	11.82	6.25	7.85	8.44	8.32	8.61	10.57	8.8	6.03	8.88

Quality scores are in the range [0,4], where 4 is the most accurate. AR, Argentina; BR, Brazil; CL, Chile; CO, Colombia; EC, Ecuador; MX, Mexico; PY, Paraguay; SV, El Salvador; UY, Uruguay; VE, Venezuela.

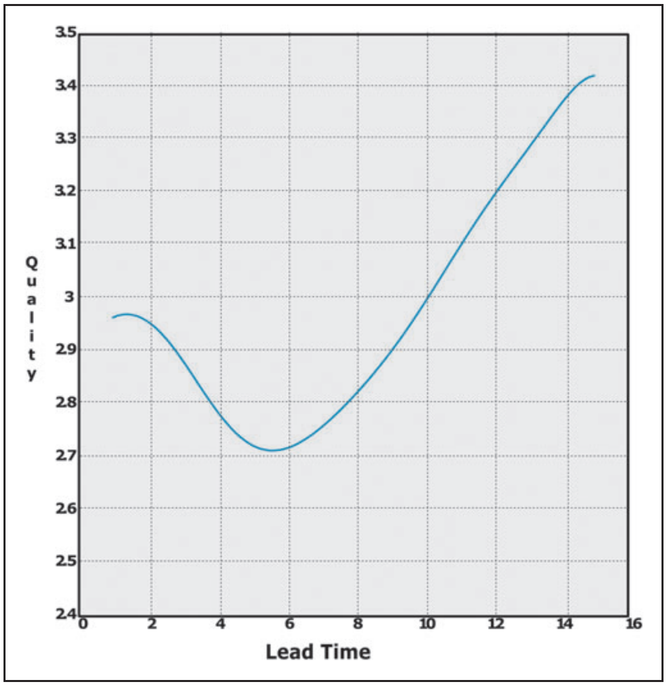


FIG. 8. Lead time versus quality score tradeoff in EMBERS.

note that there is a natural tradeoff between precision and recall, and between quality score and lead time. To explore the latter relationship, consider Figure 8, which illustrates an interesting trend. As lead time increases from low values, as expected, quality scores decrease. But as lead time crosses a threshold, quality scores actually improve again! This is because data sources like Facebook event pages and other feeds contribute high-quality planned protest warnings with high lead time.

> "AS CAN BE SEEN, EMBERS WAS QUITE ADEPT AT CAPTURING THE SIGNIFICANT ORDER-OF-MAGNITUDE INCREASE IN NUMBER OF PROTESTS, ALTHOUGH IT WAS SLOW TO FORECAST THIS UPRISING."

We now turn to specific events that were correctly forecast by EMBERS as well as some misses. The "Brazilian Spring" of June 2013 began as protests against bus fare increases but quickly morphed into broader civil unrest about wasteful government spending and lack of attention paid to multiple citizen grievances. Figure 9 illustrates the count of events in Brazil during the June 2013 uprising overlaid on the count of forecasts made by EMBERS for the same period. As can be seen, EMBERS was quite adept at capturing the significant order-of-magnitude increase in number of protests, although it was slow to forecast this uprising (the "miss" is visible in EMBERS not quite capturing the initial smaller uptick). It is important to note that the GSR for a given month is released only the following month, and thus EMBERS was able to use open-source information to forecast this large-scale uprising.

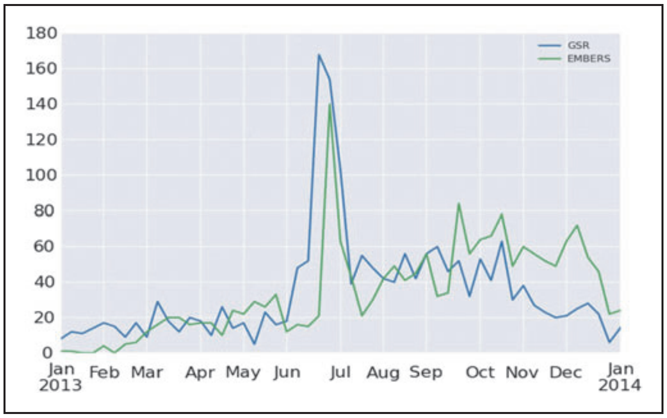


FIG. 9. Forecasting the Brazilian uprising (June 2013).

Figure 10 describes the performance of EMBERS with respect to the student-led uprisings in Venezuela (February 2014). These protests began as student disenchantment with respect to police responses and again expanded scope to larger endemic issues. As can be seen, EMBERS captures the timing of three significant upticks quite well and further (not shown in the figure) was able to categorize them as violent protests (which indeed they were). Violent protests are in general a minority (rare) class, and hence it is a significant experimental result to be able to forecast not just the protests but their propensity for violence. Furthermore, EMBERS was adept at capturing the spread of protests as they propagated to multiple cities (Fig. 11).

It is pertinent to note that the GSR, as a whole, is comprised of not just mass uprisings such as these but also a plethora of everyday protests that do not make international headlines. The combination of multiple models

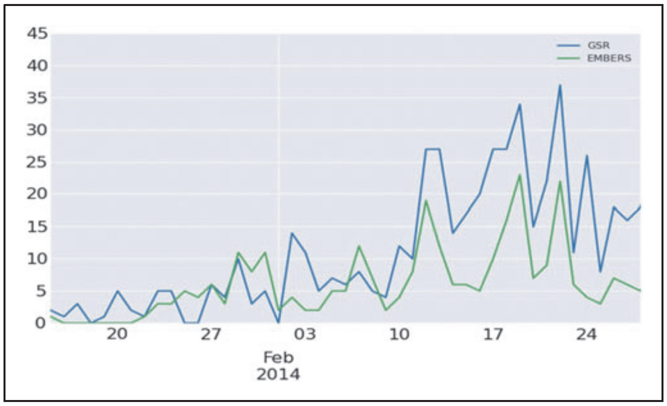


FIG. 10. Forecasting the student-led violent protests in Venezuela (February 2014).

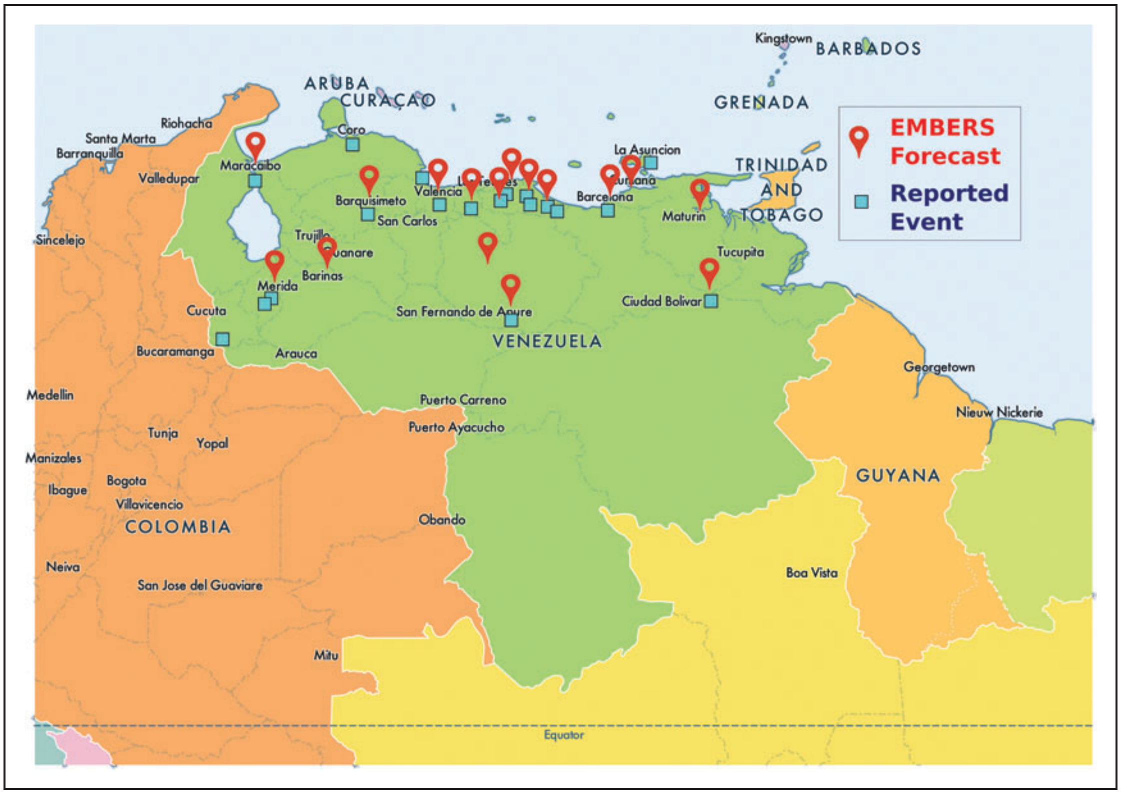


FIG. 11. Spread of the student-led protests in Venezuela across multiple cities (February 2014).

in EMBERS is adept at forecasting both classes of events as shown by the quantitative figures in Table 1 and by the mass uprisings in Brazil and Venezuela.

Discussion

A system like EMBERS has several key uses across multiple sectors of industry and government. Numerous disciplines can directly benefit from the alerts of civil unrest issued by our big data methods. For a social scientist, EMBERS helps explain how citizenry express themselves through modern means of communication. For a traveler, EMBERS can help provide alerts about hotspots and potential disruptions. For the government, EMBERS helps prioritize citizen grievances into those that are imminent in leading to protests. Various other industries such as supply chain management directly benefit from EMBERS alerts.

EMBERS thus presents a working example of a big data streaming architecture designed to process large volumes of social media data and produce predictions using a variety of modeling approaches. While EMBERS is primarily a research platform, the operational experience with the system indicates that the streaming message-based architecture is a viable approach to big data system implementation and that it performs well in some real-world scenarios that tested its ability to forecast large atypical events.

Acknowledgments

This study was supported by the IARPA via DoI/NBC contract number D12PC000337. The U.S. Government is authorized to reproduce and distribute reprints of this work for governmental purposes notwithstanding any copyright annotation thereon.

We would also like to acknowledge the contributions of a number of our team members, including Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, Chris Kuhlman, Achla Marathe, Ting Hua, and Bert Huang of Viginia Tech; Aravind Srinivasan and Khoa Trinh of University of Maryland; Lise Getoor of University of California at Santa Cruz; Feng Chen of University at Albany–State University of New York; Jaime Arredondo and David Mares of University of California at San Diego; Dipak Gupta of San Diego State University; and John Brownstein, Sumiko Mekaru, and Elaine Nsoesie of Children's Hospital Boston.

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/NBC, or the U.S. Government.

Author Disclosure Statement

All authors declare no competing financial interests exist.

References

1. Ramakrishnan N, Butler P, Muthiah S, et al. “Beating the news” with EMBERS: forecasting civil unrest using open source indicators. In: Macskassy SA, Perlich C, Leskovec J, et al. (Eds.): Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ′14, New York, ACM, August 24–27, 2014. pp. 1799–1808.
2. Bushmann F, Meunier R, Rohnert H, et al. Pattern Oriented Software Architecture: A System of Patterns. New York: John Wiley & Sons, 1996.
3. Chardonnens T, Cudré-Mauroux, P, Grund, M, et al. Big data analytics on high Velocity streams: A case study. In: Hu X, Lin TY, Raghavan V, et al. (Eds.): Proceedings of the IEEE International Conference on Big Data, Santa Clara: IEEE, October 6–9, 2013. pp. 784–787.
4. Garzó A, Benczúr AA, Sidló CI, et al. Real-time streaming mobility analytics. In: Hu X, Lin TY, Raghavan V, et al. (Eds.): Proceedings of the IEEE International Conference on Big Data, Santa Clara: IEEE, October 6–9, 2013. pp. 697–702.
5. Tan T, Ma RTB, Winslett M, et al. Resa: realtime elastic streaming analytics in the cloud. In: Ross KA, Srivastava D, Papadias D, (Eds.): Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD ′13, New York: ACM, June 22–27, 2013. pp. 1287–1288.
6. Santos I, Tilly M, Chandramouli B, et al. DiAl: distributed streaming analytics anywhere, anytime. In: Böhlen M, Koch C, Jagadish HV, et al. (Eds.): Proceedings of the VLDB Endowment, Riva del Garda, Trento, Italy: VLDB Endowment, August 26–27, 2013, 6, pp. 1386–1389.
7. Raymond ES. The Art of Unix Programming. New York: Addison-Wesley Professional, 2003.
8. Brownstein JS, Freifeld CC, Reis BY, et al. Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med 2008; 5:e151.
9. Freifeld CC, Mandl KD, Reis BY, et al. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J Am Med Inform Assoc 2008; 15:150–157.
10. Rodell M, Houser PR, Jambor UEA, et al. The global land data assimilation system. Bull Am Meteorol Soc 2004; 85:381–394.
11. Llorens H, Derczynski L, Gaizauskas RJ, et al. TIMEN: An Open Temporal Expression Normalisation Resource. In: Calzolari N, Choukri K, Declerck T, et al. (Eds.): Proceedings of 8th International Conference on Language Resources and Evaluation, LREC ′12, Istanbul, Turkey: European Language Resources Association (ELRA), May 21–27, 2012. pp. 3044–3051.
12. Leetaru K, Wang S, Padmanabhan A, et al. Mapping the Global Twitter Heartbeat: The Geography of Twitter. First Monday, 2013; 18.
13. Bröcheler M, Mihalkova L, Getoor L. Probabilistic Similarity Logic. In: Grünwald P, and Spirtes P, (Eds.): Proceedings of 26th Conference on Uncertainty in Artificial Intelligence, UAI ′10, Catalina Island: AUAI Press, July 8–11, 2010. pp. 73–82.
14. Bradley MM, Lang PJ. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1, Gainesville: The Center for Research in Psychophysiology, University of Florida, 1999.
15. Redondo J, Fraga I, Padrœn I, et al. The Spanish adaptation of ANEW (affective norms for English words). Behav Res Methods 2007; 39:600–605.

Address correspondence to:

Andy Doyle
CACI Inc.
4831 Walden Lane
Lanham, MD 20706

E-mail: adoyle@caci.com