BASIS TECHNOLOGY

www.basistech.com
info@basistech.com
617-386-2090

VirginiaTech

Discovery Analytics Center

Case Study

Forecasting the Future:
The EMBERS Predictive Analytics Success Story

RLI ROSETTE Language Identifier

RBL ROSETTE Base Linguistics

REX ROSETTE Entity Extractor

Would Scotland be an independent country today if it had not scheduled a vote? Some say the United Kingdom narrowly held onto Scotland because the British government pledged "more self-rule" to Scots less than two weeks before they were to vote on Scottish independence. Polls indicated the vote was "too close to call" for weeks leading up to the referendum on September 18, 2014.

Advance warning was the key. The UK government was able to act because they were forewarned, but it is rare that imminent, major, disrupting events come scheduled with so much lead time. Until now.

For about two years, the EMBERS project has been forecasting civil unrest events in Latin America with an average of seven days lead time. This project led by Virginia Polytechnic Institute and State University (Virginia Tech) is one of the most positive and significant examples of the much-touted power of big data living up to its reputation. That is, if properly mined and harnessed, open source big data can reveal startling insights with real-world impacts.

Open source big data can reveal startling insights with real-world impacts.

THROWING DOWN THE GAUNTLET

Following the Arab Spring—a series of populist upheavals in the Middle East from early 2011—government analysts in the Office of the Director of National Intelligence (ODNI) asked "Could we have foreseen these events?" That question became an initiative put forth by the Intelligence Advanced Research Projects Activity (IARPA) called the Open Source Indicators (OSI) Program, which challenged applicants "to develop methods for continuous, automated analysis of publicly available data in order to anticipate and/or detect significant societal events, such as political crises, humanitarian crises, mass violence, riots, mass migrations, disease outbreaks, economic instability, resource shortages, and responses to natural disasters." Essentially to "beat the news."

TAKING UP THE GAUNTLET

In April 2012, Dr. Naren Ramakrishnan, Director of the Discovery Analytics Center at Virginia Tech organized a multidisciplinary team from academia and industry to launch the EMBERS (Early Model-Based Event Recognition using Surrogates) project, with an initial focus on forecasting population-level events (civil unrest, elections, disease outbreaks, and domestic political crises) in Latin America. EMBERS was to realize the aims of the OSI Program by automating the generation of alerts so that analysts could focus on interpreting the discoveries, rather than the mechanics of integrating information.

Dr. Ramakrishnan has over 20 years experience working with big data, including his PhD work in computer sciences at Purdue University, current professorship at Virginia Tech, and leadership of its Discovery Analytics Center. The center brings together researchers from computer science, statistics, mathematics, and electrical and computer engineering to tackle knowledge discovery problems in intelligence analysis, sustainability, and electronic medical records. He is also an active contributor and reviewer for numerous academic publications, and has received many awards for teaching and research excellence.

The approach of Dr. Ramakrishnan's team was a human/computer collaboration.

Models were setup to monitor increased chatter about the shortage of essential goods. Sure enough, a shortage of toilet paper was connected to protests there.

They combined human expertise from subject matter experts (SMEs)—to devise and seed models—with computational power and natural language processing—to cope with the sheer enormity of examining open source big text.

For Venezuela, SMEs pointed out that because prices are heavily controlled by the Venezuelan government, shortages naturally arise from this structure. Therefore models were setup to monitor increased chatter about the shortage of essential goods. Sure enough, a shortage of toilet paper was connected to protests there.

As a second example, Ecuador is a popular country for referenda because the various constitutions of the past 20 years permit any laws proposed by the president and rejected by the Congress to be taken to a referendum. Thus close attention to lawmaking and popular votes was important in Ecuador. In these ways, SMEs shared very local, country-specific issues, that helped the engineers design better models.

MAY THE BEST ALGORITHM WIN

From the outset, it was clear to Dr. Ramakrishnan that EMBERS needed to be nimble, to test models quickly, and then rapidly incorporate what they learned with each iteration. Thus, instead of one master algorithm that tries to forecast everything, the team at Virginia Tech took an ensemble approach.

A master "fusion" module probabilistically combines the forecasts of the various individual algorithms into one final forecast.

Different content — medical news, tweets, political news, activist blogs — is fed to a diverse set of modules that focuses on forecasting different events: elections, civil unrest, disease outbreaks, etc. This works out to 6-8 algorithms developing alerts for each event class, with each algorithm having different biases, and using different combinations of data and models to produce competing forecasts.

In the end, a master "fusion" module probabilistically combines the forecasts of the various individual algorithms into one final forecast. Perhaps two algorithms of the six tend to be more accurate in Argentina, while some others have a better understanding of El Salvador, and the fusion module learns to weigh the predictions of these algorithms accordingly.

"The fusion engine combines different forecasts from competing modules, like a mix of expert analysts. We've found this to be one of the best ways to keep improving the system, rather than making one magic model that tries to do everything, but does nothing really well," Dr. Ramakrishnan said.

EXTRACTING SIGNALS FROM THE NOISE: THE UNSTRUCTURED DATA CHALLENGE

The obvious challenge of Big Data is the sheer quantity of "stuff" that has to be examined to find the useful bits that form a pattern or coalesce into "pictures" that support a forecast.

Imagine a beach where each grain of sand is like a mosaic tile; however, many of the tiles are tied up in bags both large (news articles) and small (tweets). In some cases the bags may be neatly identified as "square, blue, glass"—what we call "structured data"—but in other cases, it takes opening the bags and dumping out the contents to sort them by color, material, shape, or other criteria—analogous to "unstructured data."

For Latin America, at least 60% of EMBERS' alerts are generated from unstructured data: 35% from social media (including tweets) and 25% from news stories.

For Latin America, at least 60% of EMBERS' alerts are generated from unstructured data: 35% from social media (including tweets) and 25% from news stories. The remaining 40% come from a mix of sources including historical data and, highly structured data (such as food and commodity prices, economic indicators), and other reports.

MESSAGE ENRICHMENT: RESPONDING TO THE CHALLENGE

So, how does EMBERS manage to label all these tiles of information for its forecasting modules? A "message enrichment" step in EMBERS structures the unstructured data with the help of Basis Technology's Rosette® text analytics platform. Rosette is the "bag-opener and sorter," enriching the text and applying metadata to feed the next steps in the process. For example, Rosette combs through the Twitter feed, news feeds, and blogs, sorting them into categories: "Spanish, Portuguese, English, French" or "noun, verb, adjective" or "date/time, person, location, organization."

Rosette is the "bag-opener and sorter," enriching the text and applying metadata to feed the next steps in the process.

Additional enrichment modules may add yet more information to the Rosette output. For example, taking extracted dates and times, such as "el sábado próximo," (="next Saturday") and converting them to actual dates. Or passing mentions of locations to a geocoder to convert them into geographic coordinates.[1]

Not every EMBERS module uses data from Rosette, but for those that do, the enrichment is indispensable. Once each "tile" has been tagged and labeled, the modules can pick out the necessary ones from among the trillions of pieces of data.

From the start, Basis Technology engineers worked closely with the Virginia Tech team to configure Rosette, making small adjustments to accommodate the needs of the various forecasting modules.

"It was good that Basis Technology was adaptable in meeting our needs. This is an iterative process, and if something is not working, we need to adjust," said Dr. Ramakrishnan. "We made several changes to Rosette in the beginning to

[1]page 4, Ramakrishnan, Naren et al, "'Beating the News' with EMBERS: Forecasting Civil Unrest using Open Source Indicators" KDD '14 August 24-27, 2014

have it take into account the various types of data. But once we were happy with the output, it became a convenient black box for integration, supporting many different languages and many different language processing functions."

"We're proud to be part of this groundbreaking project," said Bill Ray, VP of Federal Sales at Basis Technology. "Because of the flexible nature of our Rosette linguistics platform, we've been able to adapt it to the needs of the EMBERS project, and in the process gain new insights and share best practices."

THE EMBERS SYSTEM IN ACTION

EMBERS is a fully automatic system running 24x7 without human intervention that digests nearly 20GB of open source data a day. The data comes from over 19,000 blog and news feeds, tweets, Healthmap alerts and reports, Wikipedia edits, economic indicators, opinion polls, weather data, Google Flu Trends, and even some non-traditional data sources, like parking lot imagery and online restaurant reservations.

EMBERS began operation in November 2012, focusing on 20 Latin American countries and producing "warnings" that forecast these events.

For instance, a civil unrest warning, comes with several pieces of information:

EMBERS is a fully automatic system running 24x7 without human intervention that digests nearly 20GB of open source data a day.

- **When:** predicted date of event
- **Where:** location of event to the city-level
- **Who:** population segment
- **Why:** reason for unrest
- **Probability:** confidence level of the prediction
- **Forecast date:** date the warning was produced

Flexible Phrase and Keyword Matching

Some of the warnings are generated by detecting mentions of future dates and times. Others are generated by machine learning models. In both cases, fine-grained information extraction is key.

For instance, calls for protests may be embedded in a short phrase in social media messages, containing key information about the date, time, location, and significance of an event. Therefore, knowing the role of each word in unstructured text is invaluable.

The modules that process flexible phrase and keyword matching are looking for similar messages in social media. The Rosette Language Identifier identifies the language of each message, and then Rosette Base Linguistics applies a language-specific module to tag each word of say, a tweet, with part-of-speech information. Based on that structure, EMBERS modules can match "chamar protesto" to "chamar um protesto" or "chamada para um protesto" ("call protest" to "call a protest" or "call for a protest"). Then the similarly phrased messages can be screened for dates and times. The phrase dictionary is systematically enlarged by an algorithm that screens the output from Rosette to find new vocabularies for use in future runs.

Dates and Times

It's not surprising that the ability to extract and use dates and times plays a leading role in forecasting. Some time and date information is structured, such as timestamps in some messages, but when they occur in unstructured text, these entities are found by Rosette Entity Extractor. The TIMEN module then resolves these temporal mentions to absolute values, such as turning "dentro de quince días" ("in a fortnight") into "05 de Octubre 2014."

Location, Location, Location

EMBERS searches for three types of location information [2], to identify:

(1) the source location of a given message or piece of data

(2) the topical location that is under discussion in the material, and

(3) the user location of primary affiliation of the author.

[2] page 4, Ramakrishnan, Naren et al, "'Beating the News' with EMBERS: Forecasting Civil Unrest using Open Source Indicators" KDD '14 August 24-27, 2014

Some of the location information may be structured within a tweet (e.g., Twitter geotags), but other locations are mentioned in the unstructured Twitter message, e.g. "#UnidadEnlaCalle MAÑANA protesta Jueves #16Oct a las 12:15 en la Calle Principal Briceno Iragorri, Caracas" ("#UnityInTheStreets TOMORROW protest Thursday #16Oct at 12:15 in the Main Street Iragorri Briceno, Caracas").

Even locations within the names of organizations may give hints to the source or topical location, e.g., "Malaga FC" in a tweet "Roque Santa Cruz, jugador del Malaga FC visito hoy Santa Cruz en Chile." ("Roque Santa Cruz, FC Malaga player, today visited Santa Cruz in Chile.").

Through entity extraction, locations under discussion can be discovered from the text of the tweet. Structured location data—from Twitter geotags or Twitter places and text fields in the user profile—are also added to each message during the data enrichment phase [3].

The Actors

Names of people and organizations extracted by Rosette Entity Extractor support an EMBERS module that detects mentions of key players. Human analysts compile a list of these significant people (limited to public figures) and organizations for the module. By using the results of the entity extractor, the system is more accurate than just conducting a word-to-word match.

Consider that words used for personal names can also represent things that are not people.

Key: person, location, time

"Walter Portugal, nació en Lisboa, Portugal, en 1976, con cuatro años se trasladó a Chile."

("Walter Portugal was born in Lisbon, Portugal in 1976, at the age of four he moved to Chile.")

The above sentence uses the word "Portugal" both as a person and as a location.

[3] page 4, Ramakrishnan, Naren et al, "'Beating the News' with EMBERS: Forecasting Civil Unrest using Open Source Indicators" KDD '14 August 24-27, 2014

THE PATH FROM 0 TO 50 ALERTS/DAY

To evaluate the success of forecasts from the Virginia Tech team and the two competing teams, the accuracy of warnings were judged by an independent, third-party group, MITRE. From the start MITRE was tasked to develop "ground truth" by looking at newspaper articles for reports of civil unrest. The MITRE team generated these gold standard reports (GSRs) which were used both as training data for the various team's models, and as a criteria for measuring success.

EMBERS started delivering warnings within six months (in November 2012) for Latin America, and by the end of the first year, was demonstrating some predictive power, but not enough to call it an unqualified success. The minimum quality standard as determined by the IARPA challenge was a 3.0 on a 4 point scale (with 4 being perfect). At the end of year one, EMBERS was flirting with this minimum quality score, but not exceeding it.

By the second year, EMBERS was consistently producing forecasts rated well above 3.0 with better lead times for civil unrest in particular.

According to Chris Walker, project manager of EMBERS, about 18 months into the project, new approaches to tune and optimize the generation of warnings were developed and that led to a big improvement in performance. The team developed a suppression engine that learns to estimate the quality of warnings and automatically suppresses those that are deemed to be of poor quality.

By the second year, EMBERS was consistently producing forecasts rated well above 3.0 with better lead times for civil unrest in particular. In addition to the suppression engine, the second factor for success was that the team had figured out which data sources added the most efficiency to forecasting, and how to adjust the ensemble of models to capitalize on this insight. For example, restaurant cancellations at OpenTable.com were highly linked with flu, and satellite photos showing fuller hospital parking lots were linked with disease spread.

THE FRUITS OF LABOR

By March of 2014, after 17 months of producing alerts, EMBERS was beating the news and the competition:

- Over 10,000 warnings delivered
- Around 40-50 warnings per day
- Correctly forecasted the protests during the "Brazilian Spring" in the summer of 2013, which were spread out over three weeks, involving hundreds of protests.
- Correctly forecasted student-led protests in Venezuela in early 2014. EMBERS also correctly forecasted that the Venezuelan protests would turn violent, as they did.
- EMBERS exceeded its two-year metrics goals in three criteria, met on one, and underperformed—by very little—in a fifth criterion.

By March of 2014, after 17 months of producing alerts, EMBERS was beating the news and the competition.

FUTURE FOR EMBERS: THE MIDDLE EAST AND BEYOND

In June 2014, fresh off their success in Latin America, the EMBERS team started work on forecasting events in the Middle East. This new regional focus on seven countries in the Middle East presents different challenges to both the human and computer aspects of the collaboration.

Human Challenges

Key to producing reliable alerts in the Middle East will be understanding the influence of cultural issues on forecasting. The EMBERS team is learning to model for things that have no parallels in Latin America. For example, lessons into how Latin American citizens express discontent do not quite hold for the Middle East.

"The Middle East is a big circle, and each country has a different cultural and historical context which has to be modeled, so our subject matter experts are crucial in figuring out what to look out for and model," Dr. Ramakrishnan said. "We have to understand what a protest means in a particular country, because the cultural context may imply something different in a different country."

Linguistic Challenges

The addition of Arabic has required some adjustments on the text processing components of EMBERS.

Existing geolocation tools need to have Arabic location entities in Arabic script translated to the Roman A-to-Z script. To address that, Rosette Name Indexer and Rosette Name Translator have been added to the stack of Basis Technology analytics working with EMBERS.

The accurate translation of words for geolocation is tricky as certain Arabic words and phrases have shifting meanings from one dialect to the next, and common nouns may be spelled identically to proper nouns.

One Arabic word can mean "two uncles" or "Amman" (the capital of Jordan) depending on the context.

Arabic adds a unique twist in that social media users may write in Arabizi (Arabic words written with Western A-to-Z letters and numbers standing in for Arabic characters). Before any natural language processing software can analyze Arabizi, it must be translated into standard Arabic script.

If enough Arabic social media data is in Arabizi, then Rosette can help tackle the very thorny task of converting Arabizi to standard Arabic script.

CUSTOM ENTITY EXTRACTION FOR EVERY DOMAIN AND DATA

Currently a single instance of Rosette processes a diverse number of data sources and its output is consumed by a wide range of EMBERS modules. A possible future refinement is to add multiple instances of Rosette Entity Extractor, each one custom-trained for a particular style of writing (e.g., tweet vs. news article) or for a particular subject matter domain (e.g., medical vs. political).

Just as EMBERS models get better over time, upon seeing more data, the Rosette Entity Extractor will identify entities more accurately when it has been trained on greater volumes of data. Twitter feeds with their own acronyms, slang, and shorthand are a world away from a CNN news article. Medical reports from the CDC use an entirely different vocabulary than a political blog.

Dr. Ramakrishnan has considered the idea of looking for "case counts" (number of people infected, died, at risk) medical reports and news articles about disease outbreaks, which would mean custom training Rosette Entity Extractor to recognize a more specific numerical entity than just the numbers it currently finds. "We haven't had a chance to explore this in detail yet but it's definitely a direction worth pursuing," Dr. Ramakrishnan said.

There is no predicting what new insights EMBERS will produce in the future.

References

Conference call with Naren Ramakrishnan, Director of the Discovery Analytics Center at Virginia Tech University; Chris Walker, EMBERS Project Manager; and Peter Hauck, EMBERS Data Scientist, September 17, 2014

Email from Naren Ramakrishnan September 29, 2014

Ramakrishnan, Naren et al, "'Beating the News' with EMBERS: Forecasting Civil Unrest using Open Source Indicators" KDD '14 August 24-27, 2014

About Basis Technology

Basis Technology develops innovative products and solutions incorporating multilingual text analytics and digital forensics.

The Rosette linguistics platform provides morphological analysis, entity extraction, name matching, name translation, and Arabic chat translation, yielding useful information from unstructured data in such fields as information retrieval, government intelligence, eDiscovery, and financial compliance.

The digital forensics team pioneers better, faster, and cheaper techniques to extract forensic evidence, keeping government and law enforcement ahead of the exponential growth of data storage volumes.

To learn more about Basis Technology's text analytics solutions, visit http://www.basistech.com/text-analytics/rosette

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints of this work 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/NBC, or the US Government.