BABELSTREET

CASE STUDY

Forecasting the Future: A Predictive Analytics Success Story

No doubt most politicians and government officials wish their positions came with a crystal ball to help them see future events. That kind of foresight could help leaders prepare and take appropriate steps to quell civil unrest, deploy humanitarian relief, or employ other mitigating actions.

Unfortunately, it is rare that imminent disruptive events come with enough lead time. Until the EMBERS project was created.

From 2021 to 2016, the EMBERS project forecasted 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), was one of the most positive and significant examples of the much-touted power of big data living up to its reputation. It demonstrated that if properly mined and harnessed, open source big data can reveal startling insights with real-world impacts.

The Challenge

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 information (PAI) 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.[1]

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 in Latin America, such as civil unrest, elections, disease outbreaks, and domestic political crises. EMBERS was designed 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.

The obvious challenge of big data is the sheer quantity of content that has to be examined to find the useful pieces that form a pattern and

[1] IARPA Open Source Indicators, https://www.iarpa.gov/research-programs/osi

support a forecast. This content can be neatly categorized as structured data or in unstructured form.

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

The Solution

How did EMBERS process this information? A message enrichment step in EMBERS structures the unstructured data with the help of Babel Street Text Analytics, an AI-powered text analytics platform. Text Analytics enriched the text and applied metadata to feed the next steps in the process. For example, Text Analytics combed through Twitter feeds, news feeds, and blogs, sorting them into categories: "Spanish, Portuguese, English, French" or "noun, verb, adjective" or "date/time, person, location, organization."

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

"It was good that Text Analytics 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 Text Analytics in the beginning to have it take into account

The EMBERS System in Action

EMBERS was a fully automatic system running 24x7 without human intervention that digested nearly 20GB of open source data a day. The data came 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 sociopolitical events.

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

- **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

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

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."

The Impact: From 0–50 Alerts Per 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 group, MITRE. From the start MITRE was tasked to develop "ground truth"[2] 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 teams' 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. At the end of year one, EMBERS was flirting with this minimum quality score, but not exceeding it.

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 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.

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

[2]"Ground truth" is a term used in machine learning that refers to labeled authoritative data (in this case, events that actually happened).

- 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.

Babel Street is the trusted technology partner for the world’s most advanced identity intelligence and risk operations. The Babel Street Insights platform delivers advanced AI and data analytics solutions to close the Risk-Confidence Gap.

Babel Street provides unmatched, analysis-ready data regardless of language, proactive risk identification, 360-degree insights, high-speed automation, and seamless integration into existing systems. We empower government and commercial organizations to transform high-stakes identity and risk operations into a strategic advantage.

Learn more at **babelstreet.com.**

BABELSTREET