Applications of Text and Social Media Analytics

CONCLUSIONS PAPER

Featuring:

Denise Bedford, Goodyear Professor of Knowledge Management, Kent State University

John Cassara, Former Intelligence Officer and Treasury Special Agent

Eric Hansen, Lieutenant Colonel (Retired), Military Intelligence Corps, US Army; Engineering Support Manager, SAS

Tom Sabo, Senior Solutions Architect, SAS Federal

Kimberly Nevala, Director of Business Strategies, SAS Best Practices (moderator)

Table of Contents

Introduction

Two hours after the front-runners completed the 2013 Boston Marathon, two homemade bombs exploded near the finish line, killing three people and injuring 264 others. Within days, the Federal Bureau of Investigation (FBI) released photographs and surveillance videos of two suspects.

The suspects were swiftly identified, in part through a massive, worldwide dissemination of information and photos. Social media sites Twitter, Facebook and others were all credited with helping to identify the Chechen brothers Dzhokhar and Tamerlan Tsarnaev.

Within days, one brother was killed in a shootout with police, and the other was captured and charged. Social media was lauded for its ability to aid the government – and condemned for vigilantism, misinformation and wrongfully targeting innocent parties.

Whether you praise social media for rising to the occasion or worry about the dangers of crowdsourcing a serious bombing investigation, one thing is clear: Social media is a powerful public influence. In the words of one journalist, online communities represent a "technologically fortified and massively, instantaneously connected populace." Government agencies can ignore this force, or they can capitalize on it, if they approach it wisely.

New analytic tools with graphical user interfaces enable you to tap into social media data streams and capture a flood of new information. Text and social media analytics can create meaningful insights from that data stream – as well as from other unstructured data sources that traditional business analytics and knowledge management solutions can't access.

The opportunities abound for using this unstructured data. Government agencies can use text analytics to categorize and create new knowledge from masses of digital documents. They can use social media to gauge public sentiment, identify unmet needs, influence opinion and shape policy decisions in alignment with trends. Social network analysis can identify hidden connections and influences, such as indicators of fraud and collusive crime.

How are early adopters using analytics to generate new intelligence from text and social media? What tactics, techniques and best practices should government agencies know about? How do you address emerging concerns about appropriate use and privacy of the intelligence gleaned from social media?

Those were questions of the day at the 2013 SAS Government Leadership Summit. A panel of experts, representing decades of experience in federal intelligence agencies, the military and academia, discussed "the art of the possible" with text and social media analytics.

"People of our community, many of whom knew nothing of one another prior to the events, organized at a moment's notice to mobilize information and resources to those in need. What would have previously required professional or governmental organizations days or weeks to accomplish was improvised in real time without formal coordination as people connected and collaborated using social media platforms."

Gerald C. (Gerry) Kane,
Associate Professor of Information Systems at the Carroll School of Management at Boston College, in Big Idea: Social Business Blog
April 25, 2013

Fighting Financial Crimes with Text and Social Media Analytics

Financial crime has far-reaching implications. It makes organized crime pay. It allows drug traffickers and smugglers to expand their operations. It undermines government tax revenue and the financial community in general, because it siphons vast sums of money from legal endeavors.

Fighting the tide of illicit funds is "an investigative jigsaw puzzle," said former US Treasury Special Agent John Cassara, an industry advisor to SAS. Assembling that puzzle starts with traditional sources of financial intelligence, such as 18 million pieces of financial intelligence filed with the US Treasury every year, including about a million Suspicious Activity Reports (SARs). Then add data from wire transfers, trade and customs activities, international sources, property records, immigration and travel records, loan applications, investment brokers and associations to create a richer view of a situation, suspect or organized crime ring.

"Does social media deserve to be included in this list of data sources? Absolutely," said Cassara. "Social media has become a mainstream channel for business and personal communication and commerce. If the government doesn't start concentrating on this potential information source, we're definitely losing out."

Cassara described three ways governments can fold social media into reactive, proactive and strategic analyses:

Reactive. "Suppose I'm a criminal investigator, and I have a source who says there's a narcotics trafficking organization run by John Smith who has set up ABC Corporation as a front company. After I debrief the source, the first thing I'm going to do is go back through law enforcement databases, including financial intelligence, to see if there is anything on John Smith in the ABC Corporation." Text analytics makes it possible to automatically interrogate the data captured in documents, email and other digital files.

"You can do the same thing with social media analysis," said Cassara. "Depending on the data stream I want to capture, I might not be able to look back dozens of years, or gather specifics such as date of birth or a bank account, but I hope to be able to measure sentiment, identify potentially telling behaviors, and detect associations or networks."

Proactive. "In proactive analysis, a violation has not yet occurred; you are trying to anticipate forward to understand where to investigate," said Cassara. "Years ago I was assigned to the American Embassy in Rome, and I linked up with the Italian fiscal police to help attack Italian-American organized crime – the Mafia – by looking at the flow of dirty money back and forth between the two countries. In effect, we pioneered the early use of predictive analytics by cross-referencing findings from financial databases with law enforcement databases.

"From surveillance cameras to cell phones to Facebook to Twitter to YouTube – the Boston bombing investigation relied on it all. But in the end, it was the public and their social connections that helped police crack the suspects' identities."

Carolyn Presutti,
"Multi, Social Media Play Huge Role in Solving Boston Bombing"
Voanews.com - April 26, 2013

"The same approach can be done with social media analytics. For example, I have recently been involved in a SAS project to help a Middle Eastern country examine threat finance. I came up with a taxonomy for text analysis – a list of keywords indicative of financial crime that can be monitored in social media to identify spikes. Social media is not the panacea, but when you overlay it onto other data sources, you get new insights you never had before."

Strategic. Strategic analysis gets at the bigger picture and helps guide investigative directions. For instance, using social network analysis and text mining, you might find something in the databases that points to an uptick in money-laundering activity related to ATM machines in certain ZIP codes. You can then use social media analysis or social network analysis (or entity link analysis) to learn more about what you're seeing.

"For example, a few years ago, Department of Defense monitoring of social media revealed that the Islamic Army in Iraq was using the Internet to organize and promote terrorist activities and solicit donations to fund further activities, using the Middle Eastern equivalent of PayPal," said Cassara. "Seeing this for the first time gave us a strategic picture, a new area to focus our attention. Social media enables real-time insight, so for the first time we might be able to monitor for these types of events by seeing where organizers are directing participants to send contributions by phone payments. This is real, live financial intelligence. I'm very excited about this. It's the art of the possible. We're not there yet, but we're getting there."

"Social media has become a mainstream channel for business and personal communication and commerce. If the government doesn't start concentrating on this potential information source, we're definitely losing out."

John Cassara,
Former Intelligence Officer and Treasury Special Agent

Generating Trust for Text Analytics Results

In December 2010, a Tunisian street vendor set himself on fire, an event that turned smoldering public discontent into a wave of demonstrations, protests, riots and civil wars across the Arab world: the Arab Spring. Some say social media was the key instigator of the uprisings; others say it was just an organizing tool. Either way, social media showed its power to rally the public and spawn collective action.

Was the Arab Spring also sparked in part by WikiLeaks, the controversial online organization that publishes secret information, classified media and news leaks from anonymous sources? Government agencies around the world wanted to know, and Denise Bedford, Goodyear Professor of Knowledge Management at Kent State University, wanted to offer a credible answer. To what extent did WikiLeaks influence these events, if at all?

"If we had people read WikiLeak documents and give us their opinions about whether there was any relationship, our results would have been skewed by the filters of human bias," said Bedford. "With text analytics and social media analysis technologies, we can test subjective opinions in quantifiable and objective ways.

"We began with very solid models of what people knew about the Arab Spring and what the uprising was about. We also developed profiles of what we called diplomatic language, and looked for semantic patterns that would indicate a causal connection. Our analysis suggested that there was almost nothing in the WikiLeak documents that had anything to do with the Arab Spring – no connection at all."

Can we believe that? Yes, says Bedford. "If you can objectively model language, and then have people view that model and sign off on it, you can run that model and quantify ideas – and get results you can trust."

That wasn't always true, Bedford notes. "I had been working with text analytics and artificial intelligence in my former roles at the World Bank and at Stanford University. In the work we had done up from the mid-1980s until about 2002, we weren't sure we could trust the results we were getting from text analytics. So we set about developing a methodology for modeling human knowledge into the profiles and rule sets, and validating those models before we applied them to any sort of text.

"One of my projects at the World Bank was to automate the topic classification of all documents, regardless of which repository they were in, where they were created, what type of content (Web content, formal documents, archives, human resource records, etc.). We were able to do this effectively because we had signoff from the experts at the World Bank that the topics had been correctly modeled. That advance work was a six-month effort, but it was well worth it, because then we could use the classification with confidence across the organization."

"If you can objectively model language and then have people view that model and sign off on it, you can run that model and quantify ideas – with results you can trust."

Denise Bedford,
Goodyear Professor of Information Architecture and Knowledge Management, Kent State University

Text Analytics to Understand Campaign Rhetoric

Do you ever feel that presidential campaigns are charged with hostility, emotion and unsubstantiated claims – that the rhetoric is a little crazy? Do you leave the debates feeling a vague sense of futility and pessimism? If so, there's a reason, and it can actually be quantified, said Bedford.

"At Kent State, we did an analysis of the political discourse of the 2012 presidential election. In talking to people who knew a lot about this subject, we said, 'All right, we think this discourse was highly emotional and negatively charged. We think that it was not concrete.'" Were the presidential candidates elevating the emotion, or was the media blowing it out of proportion? What would objective analysis reveal? Bedford wanted to explore that question using more than 10 or 20 keywords, since the nuance of speech is so much richer than just keywords.

Bedford started by going to psychiatric and psychology literature to find an assessment tool designed to evaluate anxiety, depression, inward- and outward-directed hostility, and quality of life in institutionalized patients. This tool served as the basis for text analysis of the language patterns of 10 presidential candidates in the primaries – and then to assess related media coverage.

If you think the media inflated the dialogue for drama and ratings, the results of text analytics will surprise you as much as they did Bedford. “It turns out, the presidential candidates – not the media – were elevating the emotional level of the discourse. The media was a bit more tempered down, if you can believe that. This finding was totally contrary to my subjective opinion. The candidates ranked very high on outward-directed hostility and depression and low on quality of life.”

What we can do with this information is up for conjecture, but it is an interesting example of merging human knowledge with machine learning to create new understanding – objective understanding, not biased by our perceptions and assumptions. This example speaks well to the growing maturity of text and social media analytics as a tool that can produce trusted results.

Social Media as a Surveillance and Early-Warning Tool

April 2009 marked the start of the “swine flu” pandemic involving a mutated version of the H1N1 virus that killed hundreds of millions of people from 1918 to 1920. Swine flu cases were first recognized in Veracruz, Mexico. Government agencies in Mexico City closed public and private facilities in an attempt to contain the spread of the virus. Those precautions came too late. An epidemic had been going on for months before it was officially recognized. Since each infected person would infect an average of 1.75 others, the disease spread globally and killed hundreds of thousands of people.

How many of those lives would have been spared if pockets of the disease and high-risk behaviors could have been identified earlier? Social media has a role here, said Tom Sabo, a SAS Senior Solutions Architect.

“Social media can serve as an early-warning system, another signal of an impending event,” said Sabo. “Back when H1N1 was an active threat, people were posting about their symptoms, their positive or negative tests, or their hospitalizations at a rate of about 1,000 tweets per hour. From that data stream we could mine events such as cases where people were throwing swine flu parties and inviting all the neighborhood kids over when one of the kids was diagnosed with swine flu – which of course is a horrible idea.”

From his experience working with federal government clients, Sabo listed other practical applications for text and social media analytics, such as monitoring the health of the economy, looking for indications of the next economic downturn, and for agencies to monitor public reaction to policies. Text analytics can help combat cybercrimes and cyberterrorism. You can do forensic analysis on reports, pull out IP addresses, spot malware and see how these things correlate across different data sources. Similarly, text analytics can correlate unstructured data from tip lines, benefit programs, financial programs and SARs. “Given a past history of which allegations and SARs led to successful investigations, we can use that knowledge to build predictive models where the text component weighs in,” said Sabo.

Five Best Practices for Text and Social Media Analytics

1. Start with the problem in mind.

"It would be tempting to just throw all your data at an analytical model and get some results," but that's not the path to success, said Sabo. "People should start with the problem for which they want to develop a solution, then assess what data they have to support that inquiry. 'What do I have in-house to explore this question?' 'What other data can I use to supplement and correlate these sources?' You'll find you can get different results when you start with domain knowledge around the problem, as opposed to starting by doing explorations on the data. Either approach is acceptable, but be sure to approach the data with a problem."

2. Acknowledge and work with language variances.

Text analytics technologies and expertise have been around for decades, but the field is constantly changing, especially as text analytics is applied to novel sources. "The use of language is different in social media than it is in formal documents and other communication channels," said Bedford. The informality and immediacy of social media encourages people to say things they wouldn't likely say in other communications.

"For that reason, social media can actually give you far more insights," said Kimberly Nevala, Director of Business Strategies for the SAS Best Practices group. "The value of social media is in getting to that tacit knowledge."

To do that and get it right, you have to be well-versed in the domain you're talking about, said Cassara. The same language can have very different connotations in different contexts. "Radical" might be a desirable quality in a design for a concept car, but not so much in a political candidate. "Killer" is a fine compliment for a prom dress or a ski resort, but not for a suspected criminal. "If you want to find out how people feel about a subject, you have to be able to organize a taxonomy around how language is used in that context, and that requires domain expertise."

3. Use aliases to create consistency from diversity.

Language varies by social media forms, from the brevity of Twitter to the multimedia convergence of Facebook, to the conversational, first-person tone of blogs. You have to account for that diversity while maintaining consistency in the analysis, said Eric Hansen, a former Lieutenant Colonel in the US Army Military Intelligence Corps.

Even within one medium, language naturally changes over time, Hansen noted. "As an analyst in Iraq, I had a difficult time finding trends more than a year long. We realized that every time a new unit came into the country, they changed the place names. Once I was able to embed that knowledge into the tool – to say, 'Place X one year is the same as Place Y the following year' – and alias for that, I was able to create those longer-term trends.

"Previously, we would have had to force each unit to call places by the same name, but there were a number of reasons, cultural and otherwise, that we didn't want to do that. Now we don't have to try to force solutions on the input side; I can alias on the output side. One of the exciting things about this is that for the first time in history, we can create a common understanding but allow for diversity to an extent that wasn't possible before."

4. Don't treat it as a set-it-and-forget-it process.

One of the challenges of social media analytics is keeping pace with ever-changing online language. It's one thing to validate a rule set against a fixed body of language, such as a well-established psychiatric assessment tool. It's quite another to keep up with the memes, acronyms and emoticons that pop up on Twitter and Facebook, some of them quite fleeting.

Taxonomies need to adapt over time, and this maintenance is only a semiautomated process. Bedford talked about her days at the World Bank, when she would read the organization's internal newspaper every morning to pick up new topics of conversation and engineer them back into the program. Emerging topics invariably require the human touch. For cases where a language or turn of phrase appears frequently, analytical tools can identify clusters and quickly equate them to other clusters. "It is definitely a hybrid process," said Sabo. "We're not looking to replace people; we're looking to supplement what human eyes and intuition can do."

"It is definitely a hybrid process. We're not looking to replace people; we're looking to supplement what human eyes and intuition can do."

Tom Sabo
Senior Solutions Architect, SAS Federal

5. Use text analytics as a complement to traditional analytical methods.

"Social media is not the silver bullet," said Babak Akhgar, Professor of Informatics at Sheffield Hallam University, speaking at another SAS business leadership event. "We must reach a point where we can integrate traditional data sources with these new and emerging information sources to create a knowledge matrix. When you combine traditional data sources, such as human intelligence from covert and overt operations, with open-source intelligence captured from new media, you can gain a powerful vantage point to see crimes as they are emerging. Then you can shift the focus from investigating what happened to preventing what is about to happen."

Closing Thoughts

> ***"After years of secretly monitoring the public, we were astounded so many people would willingly publicize where they lived, their religious and political views, an alphabetized list of all their friends, personal email addresses, phone numbers, hundreds of photos of themselves, and even status updates about what they were doing moment to moment. It is truly a dream come true for the CIA."***
>
> **From The Onion spoof, "CIA's 'Facebook' Program Dramatically Cut Agency's Costs"**

This testimony from an actor playing a CIA leader is all satire, but with a twist – it's not that far from the truth. In the process of using social media for their own purposes, people also give government and law enforcement agencies access to a huge amount of formerly personal information, now freely and publicly available.

Of course, the Privacy Act of 1974 is still in effect, with provisions that protect the personally identifiable information of citizens. The Office of Management and Budget (OMB) states that if information is collected through the use of a third-party website or application, the agency should collect only the information "necessary for the proper performance of agency functions and which has practical utility." Agencies may analyze aggregated data that is not personally identifiable, such as trends in sentiment, demographics and community attributes.

"The ability to capture data in flight in a very formal, service-driven way has surpassed our consensus on what constitutes appropriate use of that data," said Nevala. "Google yourself and you may be shocked at how much information about you is out there."

"There is ongoing and evolving debate about weighing privacy and civil libertarian issues with a department or agency's jurisdiction," said Cassara. When creating new intelligence by using social media – which the vast majority of users tend to (mistakenly) perceive as private – do the means justify the end?

When the mission at hand is to mitigate global health risks, maintain public order, improve government's service to citizens – or to intercept some misguided activist with a bomb in a backpack – it seems vital to capitalize on analytic techniques that can serve the public good.

About the Panelists

John Cassara, Former Intelligence Officer and Treasury Special Agent

John Cassara started his 26-year government career as an intelligence operative, and then moved to be Treasury Special Agent in the US Secret Service and US Customs Service, where he was involved in detecting and preventing money laundering, trade fraud and international smuggling. Cassara then worked in the Treasury Office of Terrorism Finance and Financial Intelligence. Retired from federal government, he is a sought-after speaker and a published author on financial threat analysis.

Denise Bedford, Goodyear Professor of Knowledge Management, Kent State University

Denise Bedford drives research and delivers courses on a broad range of topics in knowledge management and information management. Some of her current research interests include the knowledge economy, information economics and text analytics methodologies to support these disciplines.

Eric Hansen, Lieutenant Colonel (Retired), Military Intelligence Corps, US Army; Engineering Support Manager, SAS

Eric Hansen comes from an intelligence background, having spent 20 years in a variety of tactical military intelligence posts around the world. Before retiring from the Army, he was the Senior Military Operations Research Analyst for the Joint IED Defeat Organization, where he led a team of 30 information analysts. He then moved to ManTech (formerly MTCSC), where he provided quantitative analysis solutions for the Department of Defense and the Department of Health and Human Services. At SAS, Hansen focuses on defense intelligence special projects.

Tom Sabo, Senior Solutions Architect, SAS Federal

Tom Sabo has spent the last eight years immersed in text and social media analytics for government and public sector clients. He has worked in such broad-ranging areas as real-time analysis of the H1N1 flu epidemic, to helping clients develop social media strategies, to using text analytics to identify emerging research trends and topics of interest. Sabo was a panelist on the Institute of Medicine's standing committee on health threats resilience, which provides input on social media strategies to the US Department of Health and Human Services and the Office of Health Affairs of the Department of Homeland Security.

For More Information

For more about SAS social media and text analytics for government: sas.com/gov

Download thought leadership on social media and text analytics:

Fighting Crime Through Social Media and Social Network Analysis
sas.com/reg/wp/corp/50462

How to Use Unstructured Data to Improve Business Decisions:
UN's Global Pulse looks to social media for jobless indicators sas.com/resources/asset/un-global-pulse.pdf

View a related webinar:

Make Connections Between Disconnected Data Fragments to Reveal Hidden Threats
sas.com/reg/gen/corp/2251466

Join the SAS Text and Content Analytics Community:

Online forum: communities.sas.com/community/support-communities/text-analytics

SAS blog, The Text Frontier: blogs.sas.com/content/text-mining/The Text Frontier

Applying Business Analytics Webinar Series: Text Analytics — An Inside Perspective
sas.com/reg/web/corp/1341538

About SAS

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 65,000 sites improve performance and deliver value by making better decisions faster. Since 1976, SAS has been giving customers around the world THE POWER TO KNOW®.

SAS Institute Inc. World Headquarters +1 919 677 8000
To contact your local SAS office, please visit: **sas.com/offices**