U.S. DEPARTMENT OF ENERGY

PNNL-22171

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

State-of-the-Art of Social Media Analytics Research

ZN Gastelum
KM Whattam

January 2013

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <http://www.ntis.gov/about/form.aspx>
Online ordering: http://www.ntis.gov

This document was printed on recycled paper.
(8/2010)

PNNL-22171

State-of-the-Art of Social Media Analytics Research

ZN Gastelum
KM Whattam

January 2013

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Submitted to: Nuclear Threat Initiative Working Group III – Societal Verification

Social media consists of the tools, techniques, and technologies that use the internet to facilitate communication in an open environment. Types of social media include social networking sites (Facebook), microblogs (Twitter), blogs (Blogspot), chat (AIM), open source mapping (Wikimapia), and photo and video sharing (Flikr, Picasa, YouTube). Current research in social media analytics tends to focus around three main domains: content analysis (popular topics, and sentiment/mood), group/network analysis (define users within a group, characterize how group members interact, and identify influential users), and prediction of real-world events or characteristics. Overviews of the most recent capabilities in each area are presented below. A brief explanation of "social media as a sensor" and challenges in future social media analytics research is included at the end.

I. Social Media Analytics:

Content Analysis

Topic Identification. Topic identification is an information analytics capability that is now being applied at the larger, real-time scale of social media data. Topic identification is used to categorize popular topics of conversation on social media in general, or related to specific brands, names, keywords, or other areas of interest. In addition to research on ingesting, analyzing, and visualizing topics or themes in social media data, researchers are finding new ways to structure large social media datasets to organize, summarize, and interact with a body of documents, both at the topic/theme level, and the individual document level, and the relationships between those topics and documents.[i] PNNL capabilities in social media topic analysis include connecting to publicly available data from Twitter, blogs, social networking sites; [ii] isolating data by geographic region; analyzing trends in topics and keywords; [iii] and following and visualizing trends in real-time. [iv, v]

Sentiment Analysis. Sentiment analysis employs dictionaries and word combinations to determine sentiment (generally positive, neutral, or negative) of a conversation, a history of posts, or a chat room. [vi] New research is looking at ways to make that analysis more accurate,[vii] and also to distinguish more complex human emotion from social media, such as happy, sad, surprised, or even distressed.[viii, ix] There is also research aimed at better analyzing irony, humor, and sarcasm within social media data.[xxi] PNNL researchers have developed capabilities for analyzing and visualizing social media sentiment.[xii, xiii, xiv, xv]

Social Multimedia Analysis. New research is expanding the traditional approach of social media analysis from textual data from sources such as Twitter and Facebook to include "social multimedia" – photos, videos, maps and other "online sources of multimedia content posted in settings that foster significant individual participation and that promote community curation, discussion and re-use of content."[xvi] Social multimedia analysis sometimes focuses on content analysis, to identify multimedia related to a single event, location, or topics (such as video clips from a specific concert, or photos of popular landmarks). Other research is designed to understand online communities surrounding social multimedia and their interactions, such as subscribers to a user's YouTube updates, or those who routinely post and comment on Flickr.[xvii]

PNNL tools and capabilities facilitate understanding large collections of multimedia data, bringing together text, images, audio, and visual information in a meaningful way. Specific PNNL capabilities include sorting and searching text by topic and metadata (when published, etc.), recognizing similarities/copies of clips in videos, sorting images and videos by color, and looking at changes in word frequency in a set of documents over time.[xviii xix]

Group and Network Analysis

Identification of groups. Social media analysts are researching how to identify new or emerging groups, understand the intent of groups, and determine similarities and differences between groups. Identification of groups can include explicit relationships such as those defined by a Twitter "follower," or a Facebook "friend," but also relationships via common activity, such as commenting or posting on common online resources. Some online community research analyzes the strength or weakness of relationships between group members, and the impact of those ties on the dissemination of information among social networks. [xx] Other researchers have examined communication styles of distinct online communities, and found differences in communication styles such as use of hash tags, timing of communications (when they occur), and language use (contractions, emotive language, use of punctuation, etc.) for online communities with different members, communication goals, and topics.[xxi] PNNL capabilities in group identification include tasks such as identifying a group of academics working together in a specific field.[xxii]

Considerable recent research in social media groups has focused on identifying key influencers within a group. Influence research has included analysis of user credibility (domain knowledge) and bandwidth (reach over social media networks), [xxiii] and message diffusion via follower/audience forwarding such as "retweets" on Twitter.[xxiv] PNNL is developing research on identifying key influencers within a group.[xxv]

Relationship Characterization. Relationships between individuals can be surmised based on their communication events, such as which users have expertise, or those in a superior position to others.[xxvi, xxvii] How things are being said (language and word choice) and what is being said online may be indicative of a variety of things including the sender's relationship to the receiver. In fact, word choice is a strong indicator of the personal and social processes individuals are engaged in.[xxviii] Within online groups, language use can be identified through the use of group-based and identity-based membership claims.[xxix] PNNL research in relationship characterization has resulted in the ability to determine the relationship between sets of individuals based on their communication events (for example, determining in a relationship which individual is superior in rank to the other) and determine expertise and roles within a group based on communication.[xxx]

User Characteristics. Some research is using social media content to determine characteristics of an online community, such as the militancy of an online community.[xxxi] On a smaller scale, other research has been used to understand social media users' personality from publicly available information from Facebook profiles, such as self-description, status updates, photos, and interests.[xxxii]

Prediction

Prediction of Real-World Events. Prediction analytics seeks to utilize social media indicators to predict future, real-world events. Prediction from social media utilizes a combination of the scale of social media coverage, sentiment analysis, and influence analysis.[xxxiii] Predictive analytics using social media have been recently used, for example, to predict movie revenues,[xxxiv] outcomes of political elections,[xxxv] financial market activity,[xxxvi] popularity of a song on the Billboard weekly chart[xxxvii], and the effects of product marketing. PNNL is funding internal research and development to improve time-series forecasting and analysis for enhanced event detection.[xxxviii]

Determining Geo-Locations. Because only a portion of social media data is tagged with geo-location, researchers are developing capabilities to estimate geographic regions from unstructured, non-geo-referenced text based on natural language processing, geo-statistics, and data-driven bottom-up semantics, [xxxix] though current capabilities are still high-level, such as differentiating users in a large city versus those at a national park.

II. Social Media as a Sensor

Social data, and the corresponding capabilities to analyze the data, tends to be used either as a passive sensor for analysis associated with an event, person, or organization, or as an active sensor that engages the public to respond to specific events.

The use of social media as a passive sensor refers to the use of real-time social data for applications such as disease tracking, disaster response, reactions to a specific event (an election or attack, for example), or just generally keeping track of the pulse of a situation. The passive approach to social media analysis assumes an ill-informed public that will describe (via tweets, posts, shares, etc.) events of interest without necessarily being able to correctly identify them, or determine their significance. In the context of nonproliferation and arms control, these application types require direct access to streaming data, real-time analysis of the data, and heavy computational resources in order to handle the large amounts of data that need to be processed in real-time. PNNL is researching the potential of using social media as a sensor to supplement and support verification activities. Rather than relying on social media users to accurately detect, recognize, and report a treaty significant event in social media, the use of social media as a passive sensor derives information from more subtle signatures of users reporting unusual sights, sounds, smells, etc. Signatures derived from social media, such as the detection of unusual events, can be used to cue other, more traditional sensors and can be combined with other data sources.

As an active sensor, researchers have explored how to engage the public to provide information on a specific area of interest. Recent examples of research into social media as an active sensor are the DARPA Red Balloon Challenge[xl] and the State Department's Tag Challenge.[xli] In both events, participants were challenged to locate geographically disparate targets in a short period of time, which would preclude any one individual solving the challenge working along. The winners of the challenges (the same team from MIT in both cases) relied on social media and mainstream news media coverage to reach out to, and recruit, team members. The State Department and DARPA challenges demonstrate how crowd sourcing via social media can be used to complete complex tasks. The latest State Department contest, the Innovation in Arms Control Challenge, is seeking ideas on how crowd sourcing can support arms control

transparency.[xlii] The use of citizen-generated data was even the inspiration for a recent U.S. Department of Health and Human Services Fusion Forum.[xliii]

III. Challenges in Social Media Research

Social media faces many of the same challenges as other open source information analytics research – mainly accessing, storing and processing information; verification of sources and dealing with misinformation and deception; and fusing various types of data. In addition to those issues, social media research faces some challenges uniquely its own.

Since the source and validity of social media content are difficult to verify, some question its place in any information analysis. Additionally, analytic workflows, methods, and tools have not been designed to incorporate these dynamic, massive datasets. The cost and infrastructure required for storing and processing large amounts of social media data is something that current researchers are dealing with, along with legal considerations[xliv] for social media data storage. Despite the availability of social media data, there has been little done in the area of fusing multiple types of social media data and more traditional data into one analytic environment. Finally, the use of social media data needs to be tailored according to how people use social media, the characteristics of the information available via social media, and how that information might be meaningfully utilized to answer relevant questions.

---

[i] Chaney, Allison J.B. and David M. Blei (2012) *Visualizing Topic Models*, Presented at the Sixth International AAAI Conference on Weblogs and Social Media, 04-08 June 2012, Dublin, Ireland.

[ii] Best DM, JR Bruce, ST Dowson, OJ Love, and LR McGrath. 2012. "Web-Based Visual Analytics for Social Media." In Workshop on Social Media Visualization (SocMedVis). PNNL-SA-86200, Pacific Northwest National Laboratory, Richland, WA.

[iii] Rose SJ, DW Engel, NO Cramer, and WE Cowley. 2010. RAKE Rapid Automatic Keyword Extraction .

[iv] Ediger D, K Jiang, EJ Riedy, DA Bader, CD Corley, RM Farber, and W Reynolds. 2010. "Massive Social Network Analysis: Mining Twitter for Social Good." In 39th International Conference on Parallel Processing (ICPP 2010), September 13-16, 2010, San Diego, California, pp. 583-893. IEEE Computer Society, Los Alamitos, CA. doi:10.1109/ICPP.2010.66

[v] Gregory ML, DA Payne, D McColgin, NO Cramer, and DV Love. 2007. "Visual Analysis of Weblog Content." In International Conference on Weblogs and Social Media '07, March 26-28, 2007, Boulder, Colorado, U.S.A. International Conference on Weblogs and Social Media, Boulder, CO.

[vi] Thiel, Killian, Tobias Kotter, Michael Berthold, Rosaria Silipo, and Phil Winters (2012), Creating Useable Customer Intelligence from Social Media Data: Network Analytics meets Text Mining, KNIME, 2012.

[vii] Rafrafi, Abdelhalim, Vincent Guige, and Patrick Gallinari (2012), *Coping with Document Frequency Bias in Sentiment Classification,* Presented at the Sixth International AAAI Conference on Weblogs and Social Media, 04-08 June 2012, Dublin, Ireland.

[viii] See: De Choudhury, Munmun, Michael Gamon, and Scott Counts, *Happy, Nervous, or Surprised? Classification of Human Affective States in Social Media.* Presented at the Sixth International AAAI Conference on Weblogs and Social Media, 04-08 June 2012, Dublin, Ireland.

[ix] Brubaker, Jed R., Funda Kivran-Swaine, Lee Taber, and Gillian R . Hayes (2012), *Grief-Stricken in a Crowd: The Language of Bereavement and Distress in Social Media.* Presented at the Sixth International AAAI Conference on Weblogs and Social Media, 04-08 June 2012, Dublin, Ireland.

[x] Reyes, Antonio, Paulo Rosso, and Davide Buscaldi (2012), "From humor recognition to irony detection: The figurative language of social media." *Data & Knowledge Engineering,* Vol. 24, April 2012, pp 1-12.

[xi] Tsur, Oren, Dmity Davidov, and Ari Rappoport (2010), *ICWSM – A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews,* Presented at the Fourth International Conference on Weblogs and Social Media, Washington, D.C., 23 – 26 May 2010.

[xii] Hui PSY, and ML Gregory. 2010. "Quantifying Sentiment and Influence in Blogspaces." In SOMA 2010 - Workshop on Social Media Analytics Held in conjunction with The 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010).

[xiii] Gregory ML, N Chinchor, PD Whitney, RJ Carter, EG Hetzler, and AE Turner, II. 2006. "User-directed Sentiment Analysis: Visualizing the Affective Content of Documents ." In Proceedings of the Workshop on Sentiment and Subjectivity in Text, pp. 23-30. Association for Computational Linguistics, Morristown, NJ.
[xiv] Gregory ML, N Chinchor, PD Whitney, RJ Carter, EG Hetzler, and AE Turner. 2006. "User-directed Sentiment Analysis: Visualizing the Affective Content of Documents ." In Proceedings of the Workshop on Sentiment and Subjectivity in Text, pp. 23-30. Association for Computational Linguistics, Sydney, Australia.
[xv] PNNL (2012), *Sentiment Analysis in In-SPIRE,* http://kdi.pnnl.gov/projects/sentiment_analysis.stm
[xvi] Naaman, Mor (2012), "Social Multimedia: highlighting opportunities for search and mining of multimedia data in social multimedia applications," *Multimedia Tools and Applications,* Vol 56, Issue 1, pp. 9-34, January 2012.
[xvii] Wattenhofer, Mirjam, Roger Wattenhofer, and Zack Zhu (2012), *The You Tube Social Network,* Presented at the Sixth International AAAI Conference on Weblogs and Social Media, 04-08 June 2012, Dublin, Ireland.
[xviii] Payne, Debbie (2011), *Multimedia Analytics,* 31 March 2011, www.vacommunity.org/article32
[xix] Pacific Northwest National Laboratory, *Information Visualization,* http://vis.pnnl.gov/core_signatures.stm
[xx] Grabowicz, Przemyslaw A., Jose J. Ramasco, Esteban Moro, Josep M. Pujol, and Victor M. Eguiluz (2012), "Social Feautres of Online Netowrks: The Strength of Intermediary Ties in Online Social Media." *PLOS ONE*, Vol. 7, Issue 1, 11 January 2012.
[xxi] Paris, Cecile, Paul Thomas, and Stephen Wan (2012), *Differences in Language and Style Between Two Social Media Communities,* Presented at the Sixth International AAAI Conference on Weblogs and Social Media, 04-08 June 2012, Dublin, Ireland.
[xxii] Ediger D, K Jiang, EJ Riedy, DA Bader, CD Corley, RM Farber, and W Reynolds. 2010. "Massive Social Network Analysis: Mining Twitter for Social Good." In 39th International Conference on Parallel Processing (ICPP 2010), September 13-16, 2010, San Diego, California, pp. 583-893. IEEE Computer Society, Los Alamitos, CA.
[xxiii] Wu, Michael (2010), *The 6 Factors of Social Media Influence: Influence Analytics 1*, 15 September 2010. http://lithosphere.lithium.com/t5/Science-of-Social-blog/The-6-Factors-of-Social-Media-Influence-Influence-Analytics-1/ba-p/5708 Accessed 30 November 2012.
[xxiv] Romero, Daniel M., Wojciech Galuba, Sitaram Asur, and Bernando A. Huberman (2010), *Influence and Passivity in Social Media*, arXiv:1008.1253 06 August 2010, http://arxiv.org/abs/1008.1253, accessed 30 November 2012.
[xxv] Hui PSY, and ML Gregory. 2010. "Quantifying Sentiment and Influence in Blogspaces." In SOMA 2010 - Workshop on Social Media Analytics Held in conjunction with The 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010).
[xxvi] Gilbert, Eric (2012), *Phrases that Signal Workplace Hierarchy,* Presented at the ACM 2012 Conference on Computer Supported Cooperative Work, 11-15 February 2012, Seattle, WA.
[xxvii] Bramsen, Philip, Martha Escobar-Molano, Ami Patel, and Rafael Alonso (2011), *Extracting Social Power Relationships from Natural Language,* presented at the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, 19-24 June 2011.
[xxviii] Adali, Sibel, Fred Sisenda, and Malik Magdon-Ismail (2012), *Actions Speak as Loud as Words: Predicting Relationship from Social Behavior Data.* Presented at the 21st International Conference on the WWW, 16 – 20 April 2012, pp. 689-698.
[xxix] Burke, Moira, Robert Kraut, and Elisabeth Joyce (2009), "Membership Claims and Requests: Conversation-Level Newcomer Socialization Strategies in Online Groups," *Small Group Research* 41(1): 4-40.
[xxx] Gregory ML, LR McGrath, EB Bell, KA O'Hara, and KO Domico. 2011. "Domain Independent Knowledge Base Population From Structured and Unstructured Data Sources." In Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference, May 18-20, 2011, Palm Beach, Florida, ed. RC Murray and PM McCarthy. AAAI Press, Menlo Park, CA.
[xxxi] Gawron, Jean Mark, Dipak Gupta, Kellen Stephens, Ming-Hsiang Tsou, Brian Spitzberg, and Li An, *Using Group Membership Markers for Group Identification*, Presented at the Sixth International AAAI Conference on Weblogs and Social Media, 04-08 June 2012, Dublin, Ireland
[xxxii] Golbeck, Jennifer, Cristina Robles, and Karen Turner, *Predicting Personality with Social Media*. Presented at ACM CHI Conference on Human Factors in Computing Systems 2011. 07-12 May 2011, Vancouver, BC, Canada.
[xxxiii] Asur, Sitaram and Bernardo A. Huberman (2010), *Predicting the Future with Social Media*, 29 March 2010 arXiv:1003.5699v1, accessed 29 November 2012.
[xxxiv] Thiel, Killian, Tobias Kotter, Michael Berthold, Rosaria Silipo, and Phil Winters (2012), *Creating Usinable Customer Intelligence from Social Media Data: Network Analytics meets Text Mining*, KNIME, 2012.
[xxxv] Boutet, Antoine, Hyoungshick Kim, and Eiko YOneki (2012), *What's in Your Tweets? I Know Who You Supported in the UK 2010 General Election,* Presented at the Sixth International AAAI Conference on Weblogs and Social Media, 04-08 June 2012, Dublin, Ireland
[xxxvi] Bollen, J.; Mao, H.; and Zeng, X. 2011. "Twitter Mood Predicts the Stock Market" *Journal of Computer Science* 2(1):1–8; and Karratzadeh, Milad and Mark Coates (2012), *Weblog Analysis for Predicting Correclations in Stock Price Evolutions,* Presented at the Sixth International AAAI Conference on Weblogs and Social Media, 04-08 June 2012, Dublin, Ireland
[xxxvii] Bae, Younggue and Hongchul Lee (2012), "Sentiment Analysis of Twitter Audiences: Measuring the Positive or Negative Influence of Popular Tweeters," *Journal of the American Society for Information Science and Technology*, 63(12):2521–2535, 2012.

[xxxviii] Laboratory Directed Research and Development project, led by Courtney Corley.
[xxxix] Adams, Benjamin and Krzysztof Janowics, *On the Geo-Indicativeness of Non-Georeferenced Text,* Presented at the Sixth International AAAI Conference on Weblogs and Social Media, 04-08 June 2012, Dublin, Ireland.
[xl] DARPA, *DARPA Network Challenge,* http://archive.darpa.mil/networkchallenge/
[xli] *Tag Challenge*, www.tag-challenge.com/
[xlii] *Innovation in Arms Control Challenge: How Can the Crowd Support Arms Control Transparency Efforts?* www.innocentive.com/ar/challenge/9933144
[xliii] *Bring Your Own Data: Opportunities and Challenges in Using Citizen-Generated Data for Situational Awareness,* 14 November 2012, www.phe.gov/about/opeo/fusion/forum/Documents/BYODforum.pdf