Connecting World Events to Social Media Precursors

Bradley J. Best[1] and William G. Kennedy[2]

[1]Adaptive Cognitive Systems, LLC, Blaine, WA, USA
[2]George Mason University, Fairfax, VA, USA
bjbest@adcogsys.com, wkennedy@gmu.edu

Abstract. World events are increasingly driven by the local populations, exemplified by socio-political shifts such as the Arab Spring uprisings, which depended on communication through social media to create coordinated action. While populist uprisings have been documented for centuries, the speed with which they now spread and the tools with which they can be measured are both heavily influenced by the worldwide adoption of the Internet. Understanding and predicting these events and the trends in popular sentiment that produce them, in real-time, as they unfold across scales ranging from that of the individual to that of nations, is critical. The challenge is in reducing the enormity of available data, much of it inconsistent, incongruous, misleading, or false, into a comprehensible pattern that helps an analyst make sense of it. The problem is to produce coherence from chaos and a useful abstraction from millions of data points. We address this data challenge by connecting world events to their social media precursors, analyzing the temporal relationships between the two data sources.

Keywords: World events, GDELT, Twitter

1 Introduction

We are interested in the social psychology of the origins of world events, primarily to get at their causes and to better understand their nature and possibly to recognize their development before they become a world event. Our team's social/policy question addressed in this effort is the following: **How can world events be predicted from their social media precursors**? This question is **important because** the causes of world events lead the events themselves and so to understand world events their precursors need to be identified and analyzed. The data sets we used are the GDELT [1] as defining world events, and the Twitter data set [2] as representing social media interactions leading to the world events. One limitation of this approach is that the temporal relationship between social media and world events can be expected to vary: Social media may be a leading indicator of world events in some cases, but a trailing indicator in others. Some world events such as natural disasters are not socially grounded and thus do not have social media presence prior to the event. Conflicts, the primary focus of most world event databases may not be signaled reliably by social media unless we were privy to diplomatic communications, though the diplomatic practice of "saber rattling" often results in news headlines that precede conflict. Uprisings and protests represent the other extreme, with social media communications often predating the conflict. By analyzing the temporal relationship between social media communications and world events, it may be possible to infer the extent to which an event is driven by public opinion. Our current effort is focused on prediction of world events that have social media precursors. From the world event classification system, CAMEO [3], we selected the category of "protests" as being the primary, bottom-up social event that could be expected to rise to the level of a world event. Our basic process was to start with protests included in the GDELT database, use these to identify appropriate identifiers, and then to select Twitter data based on those and related identifiers.

2 Selection of Protests in GDELT

We used events (protests) in GDELT as a form of ground truth. The theory here is that they are the culmination of a series of "micro" events that produce something newsworthy when they become sufficiently frequent, thereby entering the public consciousness. There is a significant challenge is in formalizing such a description. Anderson and Schooler [5] provided such a description at the level of the individual, developing a framework that described the statistics of the environment, including such domains as headlines from the New York Times and frequency of email contacts. Their purpose was to show that the human memory retention and forgetting function matched the likelihood of events occurring in the natural environment. This process model, which uses logarithmic decay to model the likelihood of prior events recurring, is the model of memory activation implemented within the ACT-R cognitive architecture [4].

Our use of this model of event likelihood is intended to uncover trends that rise to the level of public awareness through the repetition of micro-events documented by social media, Twitter specifically. To recover those micro events, however, which may occur at the rate of thousands of events per hour, requires piecing together multiple social media information sources.

3 Identification of Social Media Communications

Google Trends show a *normalized* frequency graph for a trend in searches. This data can be downloaded as a CSV file, which strips the absolute frequency numbers, labeling the maximum for that term during the period in question as 100. So Google Trends can provide a macro-level vision of search terms, but provides no information regarding how that term relates to any other term in absolute volume. On the other hand, Twitter tools exist that provide the absolute volume of communication about a social identifiers, or "hashtag", but only over a short period of time (Another tool, http://hashtagify.me/, supports identification of related hashtags and their relative usages.). For example, the website hashtags.org has a free frequency analytic tool (https://www.hashtags.org/analytics/frequency/) that shows a 24 hour trend graph for a hashtag (estimated from a 1% Twitter sample). While there's no direct data download, the data point values can be inspected by hovering the mouse, so it's feasible to do this for some small number of hashtags by hand for a proof of concept. For example, #trump is currently at 3,200 tweets per hour. We can figure this for 24 hours, giving us an estimate of *absolute* frequency for the current date. This can then be tied back to the Google Trends data, giving us an estimate of absolute frequency over the preceding months and years (that is, the Twitter absolute frequency data can be used as a de-normalization factor for Google Trends). These tools allow us to connect GDELT events to social media hashtags, supporting a temporal analysis of frequency patterns between the two.

4 Connection of World Events to Social Media Communications

Our work is situated with the CIBRE intelligent agent framework [6], which allows us to model populations of agents, each having a memory retention function that behaves according to the principles described by Anderson and Schooler [5]. For example, the model of a social movement catching on may be a combination of agent-agent direct communication combined with some level of leadership in the media (media mediated event-agent communication). The CIBRE toolkit allows us to simulate the social phenomena of one agent seeing an item in the news and then discussing it with another agent, resulting in a spike in the likelihood that it will be repeated to others. Other events may be purely produced through "word of mouth" (agent-agent communication), producing a similar spike, but only within that local social network, rather than the global spike produced through media communication. Using this simulation toolkit we are able to build agents driven by the selected Twitter data and socially interacting such that we can "incubate" social phenomena and identify those events that can be expected to be picked up as world events. This allows us to step beyond the simple frequency-based descriptive statistics of these events, and produce a more nuanced dynamic interaction framework, moving us closer to understanding how and why some trends rise to the level of global events, and others fade away.

5 Discussion

This work is not yet reached maturity, but will before the conference. We expect to find the precursors in the Twitter data that lead to events, which occur in the GDELT data. Using these results and insights to populate a group of agent memories, will support social simulations tracking the strengthening of a belief/idea across time within an agent framework by tying that agent's strength of belief to social media data. We think these will be pretty interesting results.

References

1. Leetaru, K., Schrodt, P.A.: "Gdelt: Global data on events, location, and tone, 1979–2012." *ISA Annual Convention*. Vol. 2. No. 4. (2013)
2. Kwak, H., et al.: "What is Twitter, a social network or a news media?" *Proceedings of the 19th international conference on World Wide Web*. ACM, (2010)
3. Schrodt, P.A., Yilmaz, O.: "The CAMEO (conflict and mediation event observations) actor coding framework." *2008 Annual Meeting of the International Studies Association*. (2008)
4. Anderson, J. R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: "An integrated theory of the mind." *Psychological review* 111(4), 1036-1060 (2004)
5. Anderson, J. R., Schooler, L.J.: "The adaptive nature of memory." (2000).
6. Best, Bradley J.: "Inducing models of behavior from expert task performance in virtual environments." *Computational and Mathematical Organization Theory* 19(3)*, pp 370-401*. (2013)