Uncovering Mechanisms of Social Movements within the Arab Spring: A Data Mining Approach

Robert Kubinec
PhD Student
Department of Politics
rmk7xy@virginia.edu

Congyu Wu
PhD Student
Department of Systems Engineering
cw9dd@virginia.edu

February 21, 2014

Our proposed research aims to deepen our understanding of social movements during the Arab Spring through mining the wealth of textual data generated by social media users. Specifically, we propose to accomplish the following objectives:

1. Measure the human environment of the Arab Spring Twitter-sphere.
2. Develop a new methodology that can make probabilistic predictions of potential political events and social sentiment outbursts that utilizes real time and historical Twitter data.
3. Test theories concerning the rationality of social movement formation.

Background

The Arab Spring began on December 18, 2010, when the self-immolation of a roadside vendor in Tunisia launched a viral social media campaign that traveled quickly to other countries. Social media, particularly Twitter, facilitated the shaping of political debates (Howard et al., 2011), the sharing of information by all types of social actors (Lotan et al., 2011), and the organization of protests as well as the articulation of their demands (Khamis and Vaughn, 2011).

The dynamic nature of the revolution poses challenges for Middle East based international organizations to operate successfully in a culturally aware manner. Institutions ranging from military forces to embassies to businesses have sought to understand how social media can inform policymaking and lead to a culturally sensitive understanding of the region. The experience of U.S. forces in the region has shown that a lack of understanding of the cultural context can result in poorly informed decisions with adverse consequences (McFate, 2005; Holiday, 2008). For this reason, the knowledge and forecasting tools that we intend to develop will help policymakers understand the region through the perspective of the Arab street.

We define the human environment as the public sentiment and opinion of a community that may have an impact on the community's collective behavior. Anthropologists have spent decades researching human culture and behavior using survey methods, culminating in growing cultural information repositories such as HRAF (Human Relations Area File) and HTS (Human Terrain System). We want to build on these approaches by leveraging the enormous load of social media messages composed during the Arab Spring to yield real-time, quantitative measurement of the human environment.

The rapid spread of the Arab Spring has raised questions concerning the ability of protesters to make fully reasoned decisions in this chaotic environment. For this reason, we want to determine whether Arab Twitter users were more likely to rely on sources of domestic or international information before they were willing to tweet messages supporting the revolution.

Weyland (2012) argues that revolutionary waves, in which social movements spread quickly from country to country, are examples of bounded rationality because protesters privileged foreign information of protests over their domestic political context. He argues these hasty judgments led to the formation of collective action movements that faced brutal repression. However, Weyland's

assertion contradicts the long-standing theoretical approach to collective action based on Schelling (1978), who argued that people will only join a movement when they see a critical number of others doing so, thus creating a tipping point beyond which collective action is plausible (Lohmann, 1994; Kuran, 1989). These models imply that a decision to join a protest should be based on domestic information because it is the domestic regime which has the capability to inflict harm on protesters. However, if Weyland is correct, then protesters discounted domestic political information at the expense of news of international protests. For this reason, measuring domestic versus foreign information concerning political events in the Arab Spring may shed light on these important questions concerning the origin of revolutionary events.

Objectives

To measure the human environment in Twitter, we need to construct a readily accessible database of relevant political events and actors that corresponds to the spatial and temporal boundaries of the Arab Spring. This will involve the collection of a considerable amount of Twitter data as well as machine-learning content analysis of this unstructured data to achieve a coherent and usable format.

Second, once we have measured this environment, we intend to build a machine-based learning model that can determine correlations between political events and corresponding Twitter activity. This will involve developing a new methodology that can make predictions of potential political events and social sentiment outbursts utilizing real time and historical social media data. Studies have shown a general correlation between political events and social media activity, in other words, "a spike in online revolutionary conversations often preceded major events on the ground" (Howard et al., 2011). We aim to discover more specific interactions between social media and political events in order to answer important questions, such as how intense a social sentiment outburst tends to be after a potentially provocative event.

Once we have accomplished our first two objectives, we will be in a position to test hypotheses concerning social movement formation. These hypotheses are:

> H1 (bounded rationality, foreign source): Users of social media in Arab countries expressed support for revolutionary activity when they learned of protest events in other countries, causing users to make a flawed comparison between the foreign country and their own country.
>
> H2 (pure rationality, domestic source): Users of social media in Arab countries expressed support for revolutionary activity when they learned of protest events happening within their country, causing users to increase their assessment of the strength of the movement.

If hypothesis 1 is true, we expect to see more tweets concerning international political events before the protests gain support; if Hypothesis 2 is true, we expect to see more tweets concerning domestic political events before the protests gain support.

Methodology

The data set we will use is a collection of all the tweets composed between 00:00 December 1, 2010 and 24:00 March 31st, 2011. The data set contains 17,424 ten-minute time windows, and each window records the tweets of people who label themselves as living in major Middle Eastern cities as of when they post the tweet. The data set was purchased from Gnip Inc. and totals 11,777,848 tweets. The time span of the data set covers the most active period of the Arab Spring during which most vital events happened.

We propose our methodology briefly as follows. To achieve the first objective, we will employ natural language processing methods, such as topic modeling and sentiment analysis, to study the content of the tweets in order to measure individual and collective metrics of Twitter users' expression at any given time, such as topic distribution and sentiment polarity toward a given entity, whether it be a political preference or a social actor. Extant research on computational Twitter content analysis typically features straightforward queries such as sorting Twitter users by tweeting activity, but has not yet delved into deeper text mining techniques.

Topic modeling is the technique of implementing certain machine learning algorithms on a collection of textual documents in order to detect its hidden thematic structure. The latest and most commonly used school of topic modeling algorithms is Latent Dirichlet Allocation (Blei et al., 2003) and its customized extensions such as Twitter-LDA (Zhao et al., 2011)), Event-Topic-LDA (Hu et al., 2012), etc.. Sentiment analysis, or opinion mining, is the technique of identifying and extracting subjective orientations expressed in textual documents. Existing algorithms can be largely categorized into a lexicon approach and a machine learning approach, both of which have implementation tools available. We will experiment with existing methods, train our own models, evaluate the results, and also explore other relevant analytical methods to enrich the results.

To test our hypotheses, we need to create timelines of pol itical events in the Arab Spring and look into textual evidence obtained from our measuring effort. We will make use of the news articles published by major news agencies together with the GDELT (Global Database of Events, Language, and Tone) database to collect a complete series of documents that describe political events related to the Arab Spring (2013). We will then label the documents by their time, location and event nature with the help of CAMEO (Conflict and Mediation Event Observations) coding system and create a set of granular and well-annotated timelines of political events.

Once the timelines of events of different locations and types have been created, we will gauge the correlations between these timelines and the timelines of human environment metrics which have been measured earlier to examine the mutual effect that Arab Spring events of different types and Arab popular reactions have on one another. Finally, we will aggregate these correlations to construct a probabilistic forecasting engine that can provide advance warnings for potentially provocative events and sentiment outbursts and evaluate the forecasting accuracy. We will also implement and evaluate existing social conflict forecasting models such as Bayesian time series (Brandt et al., 2011), Latent Dirichlet Allocation (Schrodt, 2011), and zero-inflated count models (Bagozzi, 2011).

Bibliography

Bagozzi, B. (2011). Forecasting civil conflict with zero-inflated count models. *Manuscript. Pennsylvania State University*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Brandt, P. T., Freeman, J. R., and Schrodt, P. A. (2011). Real time, time series forecasting of inter-and intra-state political conflict. *Conflict Management and Peace Science*, 28(1):41–64.

Holiday, H. L. (2008). Improving cultural awareness in the us military. Technical report, DTIC Document.

Howard, P. N., Duffy, A., Freelon, D., Hussain, M., Mari, W., and Mazaid, M. (2011). Opening closed regimes: what was the role of social media during the arab spring? Technical report, Project on Information Technology and Political Islam.

Hu, Y., John, A., Wang, F., and Kambhampati, S. (2012). Et-lda: Joint topic modeling for aligning events and their twitter feedback. In *AAAI*.

Khamis, S. and Vaughn, K. (2011). Cyberactivism in the egyptian revolution: How civic engagement and citizen journalism tilted the balance. *Arab Media and Society*, 13(3).

Kuran, T. (1989). Sparks and prairie fires: A theory of unanticipated political revolution. *Public Choice*, 61(1):41–74.

Leetaru, K. and Schrodt, P. (2013). Gdelt: Global data on events, language, and tone, 1979-2012. In *International Studies Association Annual Conference*.

Lohmann, S. (1994). The dynamics of informational cascades. *World politics*, 47(1):42–101.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., et al. (2011). The arab spring— the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International Journal of Communication*, 5:31.

McFate, M. (2005). The military utility of understanding adversary culture. Technical report, DTIC Document.

Schelling, T. C. (1978). *Micromotives and macrobehavior*. WW Norton & Company.

Schrodt, P. A. (2011). Forecasting political conflict in asia using latent dirichlet allocation models. In *Annual Meeting of the European Political Science Association, Dublin*.

Weyland, K. (2012). The arab spring: Why the surprising similarities with the revolutionary wave of 1848? *Perspectives on Politics*, 10(04):917–934.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. pages 338–349.