Precursor to the Arab Spring, Evidence from the Social Media

Guan-Cheng Li
Electrical Engineering and Computer Sciences
University of California, Berkeley
May 7, 2012

Abstract
Using the computer to analyze a large-scale textual data from Tweet and various Blog web sites over the period of 2007 - 2012, we study the differences of turmoils, which happen to Algeria, Bahrain, Egypt, Libya, Syria, Tunisia, and Yemen. We present visualizations serve as a tool to compare different countries of Arab spring, and to study them in chronological order to track the use of sentiment words and precursors ahead of the major revolutions, which took place in February 2011.

1 Motivation
Our perceptions are largely shaped by a plethora of textual media. Counting the number of occurrences of one of the Arab Spring countries on media is one way to disclose its increasing prominence. However, to understand how the diffusion of “jasmine revolution” has resulted in the prevalence of the social media discussion today, it is imperative to devise a systematic approach that creates lists of other terms which are strongly associated with the countries across time.

2 Data
The textual data are downloaded from Lexis Nexis, at the sections of Newstex Twitter feeds and government/politics blogs, which mention one or more of these seven countries dating from January 1, 2007 to April 15, 2012. I treat each sentence as an independent sample, revert all characters to lower case and scrub all punctuation and special characters. I vectorize the text by extracting all possible uni-grams, bi-grams, and tri-grams from the sample. The aggregate corpora is comprised of 821,064 distinct terms (our dictionary) used across 2,780,050 sentences, leading to a data matrix $X \in \mathbb{R}^{m \times n}$, with $m = |D| = 2,780,050$ rows (number of samples or document units) and $n = |T| = 821,064$ columns (dimension of feature space).

3 Methods
TF-IDF re-scaling We normalize each row of $X$ to be a unit vector to prevent bias toward longer documents where higher counts do not necessarily reflect importance of that term in the document. Thus we obtain the term frequency $tf(d_i, t_j)$. To eliminate spurious features that are common across many documents such as “the” and “to”, the inverse document frequency is introduced as

$$idf(D, t_j) = \log \frac{|D|}{1 + |\{d_i \in D : t_j \in d_i\}|}$$
with $|\{d_i \in D : t_j \in d_i\}|$ the number of documents which contain term $t_j$, i.e., $\text{tf}(d_i, t_j) \neq 0$ or $x_{i,j} \neq 0$. The one added in the denominator prevents a division-by-zero in case the term is not in the corpus. Finally, $X_{\text{tf-idf}} = \{x'_{i,j}\}$ where

$$x'_{i,j} = \text{tf}(d_i, t_j) \times \text{idf}(D, t_j).$$

**LASSO** To assess the strength of associations between terms. We solve the problem

$$\min_\beta \|X_{\text{tf-idf}}\beta - y\|_2^2 + \lambda \|\beta\|_1.$$

The $l_1$–norm penalty encourages the regression coefficient vector $\beta$ to be sparse, bringing interpretability to the result. If $\lambda$ is large, then the optimal $\beta$ will be driven to be very sparse, and the LASSO model tends to allow only a few (nonzero) features that are the best predictors of the response vector, driven by the occurrences of one particular country name over documents. The measure of strength of association of other terms with a country is hence given by $\beta$ and is used across the analysis.

**Community structure** The Girvan-Newman algorithm identifies cohesive subgroups (called community detection by the authors of the algorithm) from a graph, induced by using LASSO. The procedure calculates the edge betweenness centrality of all the edges and then deletes the edge or edges with the highest value. The process will eventually increase the number of weak components, these components are the cohesive subgroups and they form a partition of the original data. Each time the number of components increases we obtain a new partition, these partitions are nested and the process continues while the number of components is less than a user specified maximum.

**PageRank** PageRank is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. PageRank can be calculated for collections of documents of any size. Here, let documents be terms, and let the directed links be strength of associations between two terms above a threshold. A probability is expressed as a numeric value between 0 and 1. A PageRank of 0.5 means there is a 50% chance that a person clicking on a random link will be directed to the document with the 0.5 PageRank. Similarly, the distribution of concepts can be found using PageRank when a person traverses from concepts to concepts (Figure 2).

## 4 Visualizations for exploratory data analysis

The rest of my report presents three visualizations, aiming to compare amongst countries, cluster countries, and analyze the breakdown of the information diffusion of certain terms by tracking their association with the countries across time.

## References


Figure 1: The upper vertical bar chart illustrates the total number of blogs/tweets which mention these seven countries, e.g., there have been 49,554 blogs/tweets mentioning “egypt”. The seven lower horizontal bar charts represent the most strongly associated other terms with the seven corresponding countries. The terms in black, in blue, in red appear exactly once, exactly twice, and more than twice across lists. In other words, the black terms are associated to only one country. Time span: Jan 1, 2007 – April 15, 2012.
Figure 2: The directed graph at the right is induced by using the seven countries (in blue) as seeds. The other terms are the predictors selected using LASSO, pointing to the query word. For example, “insubria” is a predictor for “turkey” and “algeria”. The color of each node pertains to one community, calculating using the Girvan-Newman algorithm. The left horizontal bar charts illustrate the order of PageRank scores for each of the nodes for the directed graph. The higher the PageRank score is, the more links coming to that node, both directly and indirectly.
Figure 3: This visualization analyzes the breakdown of the diffusion of certain terms by tracking their association with the seven countries across time. The upper line chart shows the number of total counts of the seven countries in blogs/tweets. It is clear to see that all these countries had a sharp up-turning point in February 2011. The lower heat map visualizes an adjacency matrix of strength of associations between a particular left-side term and the seven countries in a particular monthly time frame as indicated by the year ticks at the top. For example, the second line of the map suggests that the term democracy was initially mentioned to “syria” in Mar 2007, and recurred again in Feb and Dec 2009, and Mar 2011 onward. The last line pinpoints the fact that the concept of transition did not occur to yemen until Nov 2011. The left-side terms are differentiated by color according to each respective country. The same term is aligned vertically for temporal comparison. For example, the concept of revolution had been mentioned to “libya” and “syria” much earlier in 2007, followed by “egypt” in late 2008, and finally reached to tunisia, yemen, bahrain, and “algeria” in February 2011. The word “unrest” occurred to “libya”, “syria”, and “egypt” earlier in late 2008, followed by yemen in 2009, by “algeria” in 2010, and finally by “tunisia” and “bahrain” in February 2011.