Jaime Arredondo – JDP Student in Global Health - UC San Diego/ San Diego State University

Dipak Gupta - Department of Political Science - San Diego State University

Big data and social media messages: the discourse of social unrest in twitter.

Social Media has become an important source of big data over the last few years, particularly micro blogging tools such as twitter that allows for users to post short text based messages. The analysis of such data has led to research on social networks regarding the mobilization of actors during political protest[i], the creation of specific topic driven channels of communication[ii], and the propagation of ideas.

Lately the analysis of big data has focused on modeling and predicting future events using machine learning and data mining techniques[iii]. Previous research on social science has developed models for political instability that rely on traditional sources of statistical information, such as census and world bank data, and that span on a wide range of topics such as political, economic, demographic, and environmental information[iv]. These studies study early warning signals that forecast political instability measured by the occurrence of one of more events[v], the idea is that an automated system could perform just as good as human expert prediction. These rare events are classified according to the Integrated Data for Events Analysis (IDEA)[vi] and they are related to Open Source Indicators (OSI) reported in traditional media (TV, Newspapers).

The challenge for more accurate prediction rests upon the possibility of analyzing real time updated data sets. Twitter is a live micro blog tool that could offer the opportunity to conduct textual information on real time that reflects the day to day interest of a particular population[vii]. The prediction of civil unrest using twitter has led our research team to analyze the information for several countries in Latin America, from May 2012 until February 2013. For our experiment an automated targeted event detection system looks for terms and phrases (approximately 800) within a civic unrest dictionary[viii]; the tool looks at a random sample from daily tweet data and delivers the daily count of those words of interest.

Although a few researchers have focused on the dynamics of the social network and the spread of the message[ix], there is still a need for understanding the relationship between the set of words and the theoretical concepts of protests that lie within. We must ask ourselves what are the principal components of the discourse that lead to economic or labor protest, environmental versus electoral, constitutional order versus natural disasters.

Principal components analysis (PCA) is a statistical method that is used to discover patterns in the form of coherent subsets that are relatively independent of one another, but that include variables that are correlated with one another. These underlying relationships could be defined using previous social movements research[x] and could well vary not only between countries but also across time. Some topics might become relevant during a certain period of time or new ones might arise as information spreads and is known to the general population.

A PCA analysis makes sense if the interpretation of the observed variables (words within a dictionary) that correlate highly with each other, seem to relate under a common factor that can be tie to a theoretical concept. Within our research these factors could relate to the particular topics that comprise the discourse of social unrest in twitter. Does a protest on electoral topics must include words such as:

candidate, voting, fraud, presidential? What are the words that lead to an economic protest? Do we have patterns of text that can help us identify the most common words used by people before they protest?

The most important contribution of this experiment is related to the interpretability of the factors according to the variables and their loadings in each model. For initial experiments using only the Venezuelan Data, results seem to give us some clear understanding of the speech that is behind them. For example, the words: Election, Democracy, Results, Voting day, Left, Candidate; these concepts could be grouped together under the category of Electoral Order. This factor can reflect the interests that Venezuelans had on the electoral conflicts that the country experienced during the last year[xi].

These results are only an analysis of an exploratory exercise, the concepts behind each factor need to be analyzed in detail using the communalities within words in order to help us understand more the dynamics behind the discourse that takes place in online social media. These experiments on big data are new and the effort to categorize the text of the messages has never been done before, its results could help us identify relationship between words that could in turn feed our early warning systems of event prediction.

---

[i] Morales, A. J., J. C. Losada, and R. M. Benito. "Users structure and behavior on an online social network during a political protest." Physica A: Statistical Mechanics and its Applications (2012).

[ii] Monroy-Hernández, Andrés, et al. "Narcotweets: Social Media in Wartime." Sixth International AAAI Conference on Weblogs and Social Media. 2012.

[iii] Radinsky, Kira, Sagie Davidovich, and Shaul Markovitch. "Learning causality for news events prediction." Proceedings of the 21st international conference on World Wide Web. ACM, 2012.

[iv] Goldstone, Jack A. A global forecasting model of political instability. Political Instability Task Force, 2005.

[v] Schrodt, Philip. "Predictive Models for Political Instability." White Paper in Response to NSF SBE 2020 (2011).

[vi] Bond, Doug, et al. "Integrated data for events analysis (IDEA): An event typology for automated events data development." Journal of Peace Research 40.6 (2003): 733-745.

[vii] Naaman, Mor, Hila Becker, and Luis Gravano. "Hip and trendy: Characterizing emerging trends on Twitter." Journal of the American Society for Information Science and Technology 62.5 (2011): 902-918.

[viii] created by a set of Subject Matter Experts (SME)

[ix] González-Bailón, Sandra, et al. "The dynamics of protest recruitment through an online network." Scientific reports 1 (2011).

[x] Eckstein, Susan. Power and popular protest: Latin American social movements. Univ of California Press, 2001.

[xi] Venezuela Elections: Chavez, Capriles Begin Presidential Campaigns
http://www.huffingtonpost.com/2012/07/01/venezuela-elections-chavez-capriles_n_1641698.html