Challenges of Twitter-based Predictions of Civil Unrest in Latin America

Gizem Korkmaz
Virginia Bioinformatics Institute, Virginia Tech

Social media is believed to be responsible for facilitating critical communication often required to fuel momentum preceding the events of civil unrest. The information technology revolution has brought with it an explosion of novel data sources data such as Twitter, Facebook, news/blogs, Wikipedia that didn't previously exist or were not accessible to researchers. This paper focuses on an ongoing research project that involves tracking open social indicators such as social media and Web searches in order to predict critical societal events (protests, strikes) in targeted Latin American countries before they were reported by local news media [1]. The project requires analyzing billions of pieces of information available through social media, such as tweets, news feeds, and web queries in order to develop models that could generate a number of warnings that will be evaluated on their lead-time; the accuracy of the warning, such as the where/when/what of the alert; and the probability associated with the alert. We explore the ability of Twitter data to act as a predictive signal of civil unrest developing network-based and volume-based statistical models.

Generating time series of relevant variables from Twitter to be used in the prediction models involves multiple steps and several challenges. Our methodology in [2] is discussed here. Needless to say, the first step involves collecting the tweets of the targeted countries. One of the open research questions is which geo-location strategy to choose. One could focus on the location of the user, the location of the posted tweet or the location that is mentioned in the tweet depending on the scope of the research question. Second, in order to obtain the subset of tweets that are relevant, a keyword dictionary is developed by subject matter experts. The dictionary includes 614 civil unrest related words (such as protest, riot), 192 phrases (e.g., right to work), and country-specific actors (public figures, political parties, etc.). The tweets are filtered using the dictionary, which also includes the translations of each keyword in Spanish, Portuguese and English. Features of different languages such as accents need to be taken into account. The number of the keywords in the dictionary is kept high in order to capture the population protesting (e.g. labor, medical workers, general population) and the reason for the protest (e.g. economic, political, etc.) in addition to time and location of the protest. A small more-focused keyword list might capture the latter but not the chatter about the reasons for unrest. On the other hand, using an extensive list results in a high number of tweets that are unrelated to protests. In order to reduce the noise in the data, the tweets in which at least 3 keywords have been present are kept.

A variety of approaches are designed to build innovative surrogates from the social media data that can be used as predictors of societal events by volume-based and network-based models. In our initial volume-based model, daily counts of these protest related keywords are used [1]. We extend this model by including additional data such as social media (news/blogs), economic indicators (currency, inflation), usage of TOR (a free software that protects an individual's identity on the internet) [3], and political event databases such as ICEWS -Integrated Conflict Early Warning System- [4] and GDELT - Global Data on Events, Location and Tone- [5].

Graph-based prediction methods model the spread of information and ideologies using the concept of "cascades" in the twitter based "follower" and "retweet-mention" networks. We hypothesize that; in general, unusually large and long cascades are likely to be indicative of future events of interest, i.e., protests. We compute the total number and size of cascades, number of participants and duration (in days) of cascades, change in the number of participants and tweets, average growth rate of tweets and average growth rate of participants. For each of these features, we also compute the minimum, maximum, median, and average of the cascade size, duration, and users, as well as the average value of the 1st, 2nd, 3rd, and 4th quartile of their distribution. We employ a regression model to predict the probability of a civil unrest event in a given day by using features based on these structural properties of the activity cascades.

In all these models, sparse data matrices are obtained due to the high number of variables. We use a logistic regression model with LASSO (Least Absolute Shrinkage and Selection Operator) to select a sparse feature set, and to predict the probability of occurrence of civil unrest events in different countries [6]. The LASSO regression minimizes the error sum of squares, with a bound on the sum of the absolute values of the coefficients. It identifies the most significant variables by reducing some coefficients to zero and shrinking the value of others. The parsimonious model becomes more interpretable and the shrinking of the coefficients improves the prediction accuracy of the model by reducing the variance of the estimated values of the coefficients.

In order to train and test these models, ground truth data is needed. These datasets are compiled by an independent group comprised of social scientists and experts on Latin America. A small set of well-reputed newspapers for each country are used to identify the instances of civil unrest events, when, where, who and why of the event, i.e. the date of the event, its geographic location, the population protesting (e.g. labor, medical workers, general population) and the reason for the protest (i.e., the event type, e.g., economic, political, resource). Finally, training the models and maximizing the area under the ROC for the tested period determine a threshold for the event probability, which is used to generate warnings for each country.

The methodology allows for detection as well as prediction of country specific civil unrest events. The highlights of our results will be discussed in the presentation.

REFERENCES:

[1]. N. Ramakrishnan et al. "'Beating the News' with EMBERS: Forecasting Civil Unrest using Open Source Indicators" KDD 2014 Industrial Track.

[2]. J. Cadena, G. Korkmaz, C. Kuhlman, A. Marathe, A.Vullikanti, N. Ramakrishnan. "Forecasting Social Unrest Using Activity Cascade" (forthcoming)

[3]. Tor Project. https://www.torproject.org/

[4] S. P. O'Brien. Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. Int. Stud. Rev., 12(1):87–104, Mar. 2010.

[5] K. Leetaru and P. Schrodt. GDELT: Global data on events, location, and tone, 1979–2012. ISA Annual Convention, pages 1979–2012, 2013.

[6]. Tibshirani R. (1996). Regression Shrinkage and Selection via The LASSO. J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288.