Real-World Behavior Analysis through a Social Media Lens

Mohammad-Ali Abbasi†, Sun-Ki Chai‡, Huan Liu†, Kiran Sagoo‡

†Computer Science and Engineering, Arizona State University
‡Department of Sociology, University of Hawai'i
Ali.abbasi@asu.edu,Sunki@hawaii.edu,Huan.liu@asu.edu,sagoo@hawaii.edu

Abstract. The advent of participatory web has enabled information consumers to become information producers via social media. This phenomenon has attracted researchers of different disciplines including social scientists, political parties, and market researchers to study social media as a source of data to explain human behavior in the physical world. Could the traditional approaches of studying social behaviors such as surveys be complemented by computational studies that use massive user-generated data in social media? In this paper, using a large amount of data collected from Twitter, the blogosphere, social networks, and news sources, we perform preliminary research to investigate if human behavior in the real world can be understood by analyzing social media data. The goals of this research is twofold: (1) determining the relative effectiveness of a social media lens in analyzing and predicting real-world collective behavior, and (2) exploring the domains and situations under which social media can be a predictor for real-world's behavior. We develop a four-step model: community selection, data collection, online behavior analysis, and behavior prediction. The results of this study show that in most cases social media is a good tool for estimating attitudes and further research is needed for predicting social behavior.

1 Introduction

The advent of participatory web has created user-generated data [1], that leave massive amounts of online "clues" that can be examined to infer the attributes of the individuals who produced data. As it becomes easier and easier to create content in the virtual world, more and more data is generated in various aspects of life for studying user attitudes and behaviors. Sam Gosling in [7] reveals how his team gathers a large amount of information about people without asking any questions but only by examining the work and living places of their subjects. As we can understand people by studying their physical space and belongings, we are now able to investigate users by studying their online activities, postings, and behavior in a virtual space. This method can be a replacement for traditional data collection methods.

Among traditional social science data collection techniques, surveys or experiments are structured and active, and generating new data is an important part of the process. The researcher defines what s/he needs, designs questionnaires or experimental treatments, and collects the data based on the results

of administering them. The results provide a greater degree of control over the measurement but is expensive, time consuming, and may even be dangerous sometimes. On the other hand, studying social media, as an alternative to surveys or experiments, can be considered as an extension of passive methods of traditional social research such as field research and content analysis to observe people's attitude.

Real-world Behavior Prediction Attitudes[1] among individuals in a population can be determined in social sciences. More specifically, attitudes may be measured using established data collection techniques, such as surveys, experiments, field research, and content analysis. Alternatively, attitudes may be determined without direct measurement through models that allow them to be predicted from individual or collective structural position and/or past actions and experiences. Such approaches are often described as exogenous and endogenous analysis, respectively [5].

Online Behavior Prediction The use of World Wide Web content to predict the attitudes and behavior of individuals or groups is an issue that is increasingly tantalizing and frustrating to social and computer scientists [13]. Conversely, information available online seems to offer a gold mine of useful data - it is copious, usually publicly accessible, can be located with the aid of search engines, and often has built-in annotations in the form of meta-tags and link information. Furthermore, because such information, when public, can be downloaded by virtually anyone with an Internet connection, regardless of location, its collection generally incurs less time, expense, intrusiveness, and danger (depending on the population being studied) than traditional primary social science data collection techniques such as surveys, experiments, and field research.

On the other hand, there is not a straightforward relationship between online content and attitudes/behavior in the real-world. Although the Web is growing exceedingly fast, some sectors of populations are more likely to use it than others [9]. Furthermore, among the many interesting research issues, novel techniques are needed to explain the relationship between web content and attitudes of those producing the content, and the relationship between these attitudes and actions on the ground by the groups that the content producers represent. Neither relationship is simple to determine, and both require the implementation of innovative methodologies in order to provide the impetus to making productive use of online information as a predictor of collective behavior.

Prediction and understanding of the attitudes and behaviors of individuals and groups based on the sentiment expressed within online virtual communities is a natural area of research in the Internet era. Ginsberg et al [6] used Google search engine query data to measure concern about influenza, which in turn was used to predict influenza epidemics. They used the idea that when many from specific area are searching for influenza or topics related to it, this is a sign that there is an epidemic in that place. O'Connor et al. [10] analyzed sentiment

[1] An attitude is a hypothetical construct that represents an individual's degree of like or dislike for something.

polarity of a huge number of tweets and found a correlation of 80% with results from public opinion polls. Bollen et al [3] used Twitter data to predict trends in the stock market. They showed that one can predict general stock market trends from the overall mood expressed in a large number of tweets. In other research, Asure and Huberman [2] used Twitter data to forecast box-office revenues for movies. They showed that there is strong correlation between the amount of attention a movie has and its future revenue.

2 Methodology

The prediction of human behavior using social media has been active for years. Most of the work in this field can be classified into two categories: extraction of attitudes and prediction of behaviors. To extract attitudes, researchers mine archived or online data and map specific data patterns to specific attitudes based on the frequent patterns. Patterns are then used to predict actions and outcomes. In order to predict real-world behavior based on online behavior analysis, we follow a 4-step procedure below.

1. Select real-world communities, and find corresponding community or communities in social media;
2. Collect attitudes and online behavior from social media;
3. Analyze online behavior; and
4. Predict real-world behavior based on observed online behavior and attitudes.

Online Community Selection The initial step is selecting an online population that in some sense represents the same group as that of a real-world community. One way to do so is base selection of both the subject population and the virtual community on a particular ascriptive characteristic [11], i.e. a characteristic that is for practical purposes static, or at least difficult to change, within individuals and groups. Such characteristics include race, ancestral religion, primary language, and country/region of origin. Contemporary cross-national social science theories of ethnicity generally accept that ethnic groups are defined by one or more ascriptive boundaries [4]. Selecting populations based on ethnic characteristics is helpful because of the stability of what individuals possess them. This in turn allows the researcher to ensure that any group comprising those who share those characteristics will be a relatively stable one whose members are relatively easy to identify and are not constantly moving in and out of the group. Our ethnic group in this study is countries involving in Arab Spring revolutions.

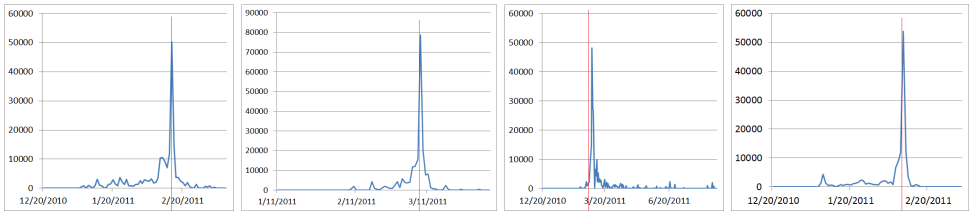
Data Collection We used Twitter to collect 35 million tweets related to Arab Spring event almost for all of the countries involved in the revolutions. In addition we collected more than one million articles and blogposts from popular Middle Eastern social tagging websites. Moreover we crawled 135,000 popular Facebook pages to collect data on posts, comments and like behavior on Facebook. The data on real-world events has been collected from Reuters.com website, which contains an archive of all published articles dating back several years.

Text processing and Online Behavior Analysis We performed preliminary text mining techniques including removing stop words, stemming, and extracting the most frequent words and phrases. Then we translated the most repeated words into English. Detecting events from newswire stories is a form of event detection from text streams, something that has been an active research field in recent computational studies of the Internet [8], [12]. In order to examine the relationship between social media artifacts and real world events, we ordered the events chronologically and performed multiple forms of statistical analysis against time-stamped results from online content analysis. For our main analysis, we used weekly word frequency data drawn from the blogosphere and compared this against a measure of the magnitude of collection action events, also weekly, drawn from the Reuters database.

Real-world Behavior Prediction We used correlational analysis to identify word categories whose frequency of mention was most significantly related statistically to the magnitude of social action during the same period, then used multivariate regression analysis to assign coefficients to them for predictive analysis. We found that our ability to predict events was preserved even when independent variables were lagged to a period of up to two weeks. For zero, one, and two week lags, we could identify at least two word categories whose coefficients were significant at the $p < 0.5$ level in a two-tailed test. Lagging independent variables has the added advantage of demonstrating that large social action events can be predicted well before their actual occurrence. Indeed, there were two categories with significant coefficients even with a lag period of 1.5 months. The number of word categories was deliberately chosen to be relatively small in number in order to ensure that significance would not be the result of spurious results due to excessive statistical degrees of freedom.

3 Observations

Event Prediction During social movements, social media has been used for either to organize future events or to report past events. For the former case we are able to find online conversation on events, even weeks before the real-world event but for the later one, as there is no or few information, prediction is not possible. We extracted frequent keywords for each event before and after that event which shows amount of online discussion for each event. Figures 1a through 1d show the amount of related discussion in social media on each topic. The peaks align very closely with the dates that real-world action activity peaked as well. This synchronization supports the hypothesis that online dialog is closely related to events on the ground. For some (usually unexpected) events, there is nothing or few conversation before the events but much more after that (Figure 1c). However there are much more tweet or blogpost after the event. The prediction results usually does not contain information about details of corresponding real-world event such as time, location, magnitude, length, effects, and participants. Also we are not able to observe or predict small events (usually those that are limited to small portion of the target society). In some events, in spite of large



(a) (b) (c) (d)

Fig. 1: (a, b, c, and d) The correlation between online behavior and real events. The vertical red line shows the day that real-world event is happened. Each graph is related to an event. x-axis is the date and y-axis is frequency of keywords related to the event. (a) Shows lots of online conversation almost 5 weeks before and two weeks after the event. This kind of events are easily predictable. (b) Almost the same as (a) but with few conversation after the real-world event. (c) shows little conversation before the event (this is the pattern for unexpected events) and much more after the events even for three months. (d) shows an event with only few days conversation after the event.

amount of online conversation, the real-world event was not as large as expected and vice versa (A large real-world event with negligible online conversation).

Attitude Extraction By mining the frequent patterns we are able to extract attitudes. For example Arab Spring tweets show that Yemenis were more concerned about Security whereas Egyptians were more concerned about Revolution and Freedom (Figures 2b and 2c). More interesting, we are able to track change of attitudes during a social movement by using a time series of tweets and blogposts.

Key People Detection By employing the same method as we used for attitude extraction, we are able to find Key people. Our definition of Key people is those whom most mentioned in social media during an event. These people usually play an important rule and knowing them in social movements is very important. The data extracted using this method is highly correlated with the data from real-world.

Mood Analysis We used method introduced by [14] to evaluate sentiment orientation of words, sentences and documents. By using this method we are able to measure the overall sentiment of the society before, during and after events. Results show that usually, people are much happier before than after events. They also are more exited when event is happening in the real-world.

4 Discussion

This section provides analyis on the results of previous section. Analysis are given for all parts of the predifined method including, Community selection, Data collection, Online behavior analysis and Prediction. In summary, the observations

show that in many cases by using social media data we are able to extract attitudes, mood, events, and even key figures. In general as we can see social media data can be used for mining purpose. But for prediction, the results are not satisfactory. In some cases resutls match with real-world data and in some cased does not. In this section we provide more details about the prediction process using social media data.

Community Selection The first step of a good prediction is attitude extraction from desired group. As we want to use social media data to extract attitudes, we should find an online substitute for our real-world group. But in most cases finding the same population both in the real-world and social media is very challenging or even impossible. This problem leads us to select a community with lower similarity with real-world group. As we can see in opinion polling, the main challenge is selecting a sample of elements from a target population. As an example, analyzing Arab Spring tweets show that roughly 75 percent of the 1 million clicks on Libya-related tweets and 89 percent of the 3 million clicks for Egypt-related Tweets came from outside of the Arab world[2]. Since almost 90% of this data is generated outside of the Arab world, how it can represent Arabs' attitudes? And in what extent the prediction based on this data is accurate?

Another example is analyzing the relation between number of followers on Twitter or number of page likes on Facebook and number of supporter in real-world. What is the relation between candidates' number of followers in Twitter and their chance to win the next presidential election? Barack Obama, Mitt Romney, and Rick Perry have 11,045,000 and 161,000 and 103,000 followers on Twitter respectively. This data is not valid for prediction because voters are not this society of Twitter user.

Data Collection Most of Twitter users who tweeted from inside of Arab world, have missing information in their profiles. Usually they don't use Twitter's geo-tag feature when they tweet. They reveal little information about themselves and usually use nicknames instead of their real name. This is very common in non-democrat countries. Lack of information about the source of data, make it unreliable and we can't use them in our analysis with confident. More seriously, people in non-democratic countries do not reveal their real believes and intension. Unrelated and spam data is another challenge in social media. Paid bloggers (or Twitterers) are another problem that misleads the prediction results. In most of the events we observed many accounts actively tweet fan of the government. Statistical methods are very sensitive to this kind of tweets.

Online Behavior Analysis Extracting attitudes from plain text due to language complexity is not a straight forward process. In most cases researchers are not able to extract complex patterns from the text. We also need to infer collective behavior by observing bunch of individuals' online behavior that generates more complexity. In some cases though there is connection between attitudes and behaviors, but it is hard to be discovered by using automatic methods.

[2] http://www.stripes.com/blogs/stripes-central/stripes-central-1.8040/researchers-skeptical-dod-can-use-social-media-to-predict-future-conflict-1.155296

(a) (b) (c)

Fig. 2: (a) The day after Steve Jobs' death his Facebook page had one million likes three times more than the day before. At the same time Apple's stock was experiencing a drop in real-world. (b, c) Tag-clouds show that Yemenis were more concerned about Security whereas Revolution for Egyptians.

Real-world Behavior Prediction Prediction is hard even if we have enough sources of data. In many cases there is not a straight forward relation between online and Real-world behavior and even by using sophisticated methods to extract online behaviors, we will not be able to predict the real-world behavior. An example of this case is events related to Steve jobs death. Before the announcement of his death, his Facebook page had less than 300,000 likes. But few days after that, more that a million of his fans liked his page (Figure 2a). People used social media and weblogs to write comments about him and his company, Apple. People everywhere praised him and his company. By using the method proposed by Bollen et al [3], we should have a huge rise for Apple's stock but in real world Apple's stock was experiencing a drop. This example and many others show that even we collect enough data from social media and run sophisticated algorithm to analyze social media data, we would not be able to predict the real-world's outcome.

5 Conclusions and Future Work

In this paper we report a comprehensive research on using social media data to analyze online behavior and predict real-world behavior. The prediction task has been divided into four sub tasks. These tasks are community selection, data collection, online behavior analysis, and real-world behavior prediction. The main challenges of this process are task one and task four which are how to identify key individuals or groups, as well as how to predict real-world behavior by interpreting the attributes of these individuals or groups. Once these two issues are addressed, analyzing the data generated by people in social media can help to understand their attitudes and to predict future real-world activities. In some cases we are not able to find a match for the real-world population, this problem causes a set of non relevant attitude and therefore misleading to wrong behavior prediction. Insufficient data in some cases is another source of failure. Compexity of language also leads to misinterpretation of online behavior. And the last item it compexity of behaviors. Sometimes information available in social media is not enough for prediction. As we see in the case of Steve Jobs' death, information available in social media is not enough to predict Apple's stock so we need more

data and need to use more comlex algorithms for prediction. We also observed that we are not able to predict minor events or details about major events.

Future research would look at comparison between these methods and more traditional social science methods, including surveys, and methods such as prediction through change in social structural variables (political and economic). Comparisons with hybrid methods such as social science attitudinal content analysis of online social media would also yield additional insights into the relationship between activity in virtual communities and on the ground.

Acknowledgments
This research is sponsored, in part, by Air Force Office of Scientific Research.

References

1. N. Agarwal, H. Liu, L. Tang, and P. Yu. Identifying the influential bloggers in a community. In *Proceedings of the international conference on Web search and web data mining*, pages 207–218. ACM, 2008.
2. S. Asur and B. Huberman. Predicting the future with social media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 492–499. IEEE, 2010.
3. J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2011.
4. S. Chai. A theory of ethnic group boundaries. *Nations and Nationalism*, 2(2):281–307, 1996.
5. S. Chai. *Choosing an identity: A general model of preference and belief formation*. Univ of Michigan Pr, 2001.
6. J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.
7. S. Gosling, D. Drummond, and I. NetLibrary. *Snoop: What your stuff says about you*. BBC Audiobooks America, 2008.
8. J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.
9. P. Norris. *Digital divide: Civic engagement, information poverty, and the Internet worldwide*. Cambridge Univ Pr, 2001.
10. B. OConnor, R. Balasubramanyan, B. Routledge, and N. Smith. From Tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 122–129, 2010.
11. T. Parsons, E. Shils, and N. Smelser. *Toward a general theory of action: Theoretical foundations for the social sciences*. Transaction Pub, 2001.
12. T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from flickr tags. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieva*, 2007.
13. M. Smith and P. Kollock. *Communities in cyberspace*. Psychology Press, 1999.
14. P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.