What Social Media Sentiment Can (And Cannot) Tell Us About Public Opinion

it's like everyday there something new mahn. Everyday. bombing. Fighting. Killing #BringBackOurGirls
Nigerian Twitter user, April 30, 2014 at 20:03 WAT.

Social media – Twitter, Facebook, blogs, and other sites where people publically share their own content – has seen enormous growth all over the world. These public traces of human activity are available in real time, in high volume, and at high data rates. They represent a source of social/cultural data at a scale not previously seen. Among other things, these data provide sources of un-mediated opinions and reactions to events, as the recent tweet from Nigeria quoted above illustrates. How can these data be utilized to obtain public opinion results similar to what is available from standard survey techniques? Surveying public opinion via social media is likely to be very different from phone surveys, face-to-face interviews, and focus groups. In social media, for example, it is difficult to obtain a representative sample of the population since the demographics and home locations of users are not always known, and if known, the sample may be highly skewed toward certain demographic groups or regions. The opinions expressed online also exhibit self-selection bias since we only hear from people who wish to share their thoughts with the rest of the world, and we only hear about the topics that they choose to talk about. These potential sources of bias in social media opinion contrast with established methods for obtaining representative samples of a population and the use of structured interviews. An understanding of these sources of bias is essential for an accurate interpretation of the opinions expressed by people via these new channels of discourse.

This talk will describe the work we have done in comparing opinion from social media with traditional public opinion surveys. Our focus is on Nigeria and opinion toward the current president, Goodluck Jonathan. We compare predictions of opinion (or in computer science jargon, sentiment) toward Jonathan in Twitter statuses (i.e. tweets) from Nigeria with opinion surveys from 2010 through 2013 for all 37 Nigerian states. We found that Twitter sentiment is consistently more negative than the levels of support and approval in the surveys, though we see strong correlations between the two data sources over time and by region. Reweighting the surveys based on factors associated with Twitter use made little difference in the comparison, suggesting that the negativity bias in Twitter sentiment was more due to self selection-bias rather than demographic or regional biases in the online population. We also see distinct differences in Twitter sentiment with respect to how and when people are tweeting. Tweets that represent direct communication from one user to another are more negative than other tweets, and the temporal context of tweets has an effect on their aggregate levels of sentiment. For example, tweets on Election Day in April 2011 showed a much higher level of support for Jonathan than on other days since people were simply using Twitter to announce their choice at the polls.

We will discuss the methods used for collecting and processing tweets, predicting sentiment, and comparing these predictions to survey results. The talk will also touch on the differences between these data sources, what we are actually measuring when we look at social media opinion, and argue how these new novel data sources should be seen more as a complement rather than a replacement for standard approaches for gauging public opinion.

Clay Fink is a senior software engineer at the Johns Hopkins University Applied Physics Laboratory. He has a B.S. in Computer Science from the University of Kentucky and an M.S. in Computer Science from the Johns Hopkins University. His work focuses on extracting sentiment and user attributes from social media text; methods for comparing such inferred aspects of online activity with ground truth data; and a range of other text processing work relating to social media and other domains. He is currently the principal investigator for the ONR funded grant "Evaluating Social Media as a Source of Public Opinion in the Developing World."