A Regional Analysis of Twitter as an Effective Means to Monitor the Spread of the 2012-2013 Influenza Epidemic

Farrell, Luke
Jacobs, Sarah

Our project aims to show the effectiveness of mass data keyword frequency over social-networking sites, such as Twitter, at accurately displaying real world occurrences and enumerating a population, accurately enough to serve as a plausible means for research. We divided the United States into 9 regions based on the 2010 population census and then proceeded to collect the exact number of tweets containing the keyword “Flu” sourced from all 9 regions, and data released by the CDC for the reported number of Influenza Like Illnesses for every week, from each region. We collected data in late May of 2013 for the 40th week in 2012 to the 22nd week of 2013, the official flu season. We then ran region by region, statistical analyses comparing keyword frequency of tweets with the reported number of influenza cases, utilizing second order polynomial regression as well as chi squared goodness of fit tests. From these tests we found that our data yielded strong correlation coefficients for each region that were consistently upwards of .9. We then began drawing conclusions about the spread of the disease throughout the United States, finding that with the systematic elimination of weeks effected by lurking variables, one can create an effective model of the number of people infected with the influenza virus based solely off of the number of tweets containing the keyword “flu”. Ultimately through our variety of statistical analyses we provided evidence supporting the idea that mass monitoring of Twitter can serve as an effective projection of popularly discussed real world events.