**NAHUSH N RAICHURA**

**THOUGHT ESSAYS**

**THE PARABLE OF GOOGLE FLU TRAPS IN BIG DATA ANALYSIS**

The reading reflects how GFT made the headlines in Feb-2013 for overestimating the proportion of doctor visits for influenza-like illness. The goal of the article is to emphasize that big data may not necessarily offer all the information, sometimes traditional small data also offers us information of relevance that cannot be obtained from big data.

The weakness of GFT was that it tried to find the best match among 50 million search terms to fit 1152 data points, the odds of finding terms matching the propensity of the flu which was structurally unrelated were high, due to which GFT predicted almost twice the results given by CDC. In spite of modifications in the GFT algorithm, the predictions made were unsatisfactory and that even a 2-week lag CDC data gave better results.

However, the actual problem to GFT was not its design but algorithm dynamics of Google’s search algorithm, since changes in the search algorithm gave erroneous correlation results for GFT. Thus, changes in the data generating algorithm must be monitored carefully so as to not affect the other systems.

The weakness of the article is, just to increase the advertisement revenue and to give better search results to its customers, Google unknowingly compromised the GFT results. However, the strength is in its solutions: how GFT and CDC can be calibrated together giving better and accurate results than GFT and CDC standalone and also how traditional small data is useful and should be considered for data analysis.

**MORE TWEETS MORE VOTES: SOCIAL MEDIA AS A QUANTITATIVE INDICATOR OF POLITICAL BEHAVIOUR**

The purpose of this article is how social media can help a candidate gain votes in the upcoming elections. The study proves how a trending candidate gains more votes independent of the user’s location and emotions and twitter data obtained is more significant than CNN data.

The strength of this article is it not only collects data from Twitter and performs data analysis, but also takes into account the variables or factors such as district partisanship, median age, percent white, etc. The model constructed shows how tweets directly affect the increase in the voters and which category of people voted the most for a candidate.

Weakness of the article found: the authors did not consider some of the data since these districts did not have any competition and they were relying on past elections results. This data should not be ignored as it could be giving some insights and it may not be necessary that a particular district may have the same result, considering some external factors might influence the outcome of the election.  Furthermore, research can be conducted on why location and emotions of the users don’t matter.

The question raised is: What if the candidate is trending for purposes other than elections (e.g. scams, non-political deals)? Or a celebrity standing for elections and the buzz on twitter is about their performances or movies, what about those tweets? There has to be a mechanism which filters out non-political tweets as the results might create a bias in the model.

**REFERENCE:**

**For thought essay 1:**

Given article of “THE PARABLE OF GOOGLE FLU TRAPS IN BIG DATA ANALYSIS” by David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani

**For thought essay 2:**

Given the article of “MORE TWEETS MORE VOTES: SOCIAL MEDIA AS A QUANTITATIVE INDICATOR OF POLITICAL BEHAVIOUR” by Joseph DiGrazia, Karissa McKelvey, Johan Bollen, Fabio Rojas.