Addressing the Demographic Bias on Twitter

Soroush Vosoughi, Russell Stevens, and Deb Roy

Laboratory for Social Machines
MIT Media Lab, Cambridge MA, USA

soroush@mit.edu

Abstract

In recent years, the ubiquity, ease-of-use and accessibility of social media have made it one of the main sources of data about human behavior. Specifically, Twitter, given its public nature, has become one of the main sources of data used to study human behavioral and social dynamics. But as is the case with any information channel, there are inherent demographic biases with Twitter. Given the thousands of scientific articles using Twitter data, one would expect this bias problem to have already been addressed. This is surprisingly not the case.

In this paper, we propose a method for correcting the distorted demographics on Twitter. We start by automatically detecting the position of a set of Twitter users towards several political issues such as abortion, gun rights, etc. This is done through a state-of-the-art "stance" system that we have developed. We then look for the distribution of these political positions in polls conducted in the real world – specifically, distribution across different demographics (e.g., what percent of males aged 18-25 believe in made global warming).

This process leaves us with three pieces of information – the distribution of political positions on Twitter, the distribution of these same positions in the real-world, and the demographic breakdown of the people polled in the real-world. We can then use a Bayesian approach to infer the demographic of the users on Twitter. Additionally, any other contextual data that is made available to us through Twitter, such as the geo-location of their users, can be used to set the priors for the Bayesian model, thus improving its performance. The model can be evaluated by using Twitter's polling capabilities, that has been made available to the public in recent months. Through this tool, we can directly collect demographic information from a few hundred Twitter users that can be used as our ground-truth for evaluation.

Keywords
Twitter, bias, demographic, Bayesian