Deconstruction Data Science

Anubhav Gupta - part 1

I have analyzed paper on “Classifying Latent User Attributes in Twitter” Delip Rao, David Yarowsky, Abhishek Sheroots, Manaswi Gupta and discussed how it stands up against the nine types of validity outlined in Krippendorff.

Face validity - Face validity means acceptance of research findings because they make sense and look believable. The paper investigates a SVM based classification algorithm to classify latent user attributes like age, gender, regional origin, and political orientation based on a rich set of features derived from the Twitter user content. This does make sense on face that the tweets or the content of the tweets of a user could be use to identify age, gender, regional origin and political orientation.

Sampling validity - For each of the attributes, the dataset was created by manually crawling the twitter network and then coming up with a seed set of users and getting further users from the follower network. Its clearly mentioned that “To avoid a la- bel bias problem we constrained the number of users in each class to be similar”[1] thus establishing sampling validity.

Functional validity - The models that are a investigated in this paper do reveal some promising results when it comes to predicting the user attributes from features derived from user text. “Our models, singly and in ensemble, significantly outperform baseline models in all cases.” thus establishing Functional validity.

---

[1] ‘Classifying Latent User Attributes in Twitter’ by Delip Rao, David Yarowsky, Abhishek Sheroots, Manaswi Gupta