CMSC 773 Final Project Proposal

Arthita Ghosh, Khanh Nguyen, Sayantan Sarkar and Varun Kumar

April 11, 2016

1 Exploratory Data Analysis

For the exploratory part we wish to explore some possible features and visualize relationships among them. Features can be of 2 types: per tweet features and per user features, both of which we can explore through the following approaches.

1. Develop a unified function that takes in groups of features from the both sets (control and schizophrenic) and then returns metrics (eg: chi square, heatmap visualization etc) to determine how important the feature is.

2. Study the changes in a single user in the per tweet features over time. For example perhaps we can find some patterns on how the features evolve over time. We may find a significant difference in some features before and after the onset of schizophrenia.

3. Apply PCA to some of the low dimensional features and plot them in 2D or 3D and explore if we can observe clusters (manually or through K-NN).

Below is a list of features that we wish to explore through the above mentioned approaches.

1.1 Features based on Twitter specific interactions

This section describes some possible features that considers the non-linguistic properties of the tweet [1]

1. Engagement: Volume of posts, Proportion of reply posts, fraction of retweets, proportion of links shared, insomnia index (time of the day tweets were made)

2. Social graph: Construct a graph of user and the people he/she interacts with and then consider the graph features like node, dyadic and network properties

The rest of the stuff in this paper is analysis of tweet sentiment etc which you guys are writing (LIWC, tree based complexity etc).

...

References

[1] M. D. Choudhury, S. Counts, E. Horvitz, and M. Gamon, "Predicting depression via social media." AAAI, July 2013. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=192721