CMSC 773 Final Project Proposal

Khanh Nguyen, Sayantan Sarkar, Varun Kumar and Arthita Ghosh

April 23, 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).

1.1 Feature extraction

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

1.1.1 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 re-tweets, 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

1.1.2 Features based on POS tags

1. POS-Tag sequence cross entropy : build a bigram model over POS tags and then measure the cross entropy of each tweet.
2. Propositional Density : number of propositions / number of words
   propositions = verbs, adjectives, prepositions and conjunctions
   main verb and all its arguments = 1 proposition
3. Content Density (POS tag based):
   open class words : NN,VB,JJ,RB or SYM
   closed class words : Prepositions
   Content Density = ratio of open vs. closed class

1.1.3 Features based on Parse Tree

1. Yngve Scoring: Size of Pushdown stack at each word in a top down left to right parser.
   Calculate size of stack at each word : at each node, score the branches from that node to its children. Start with a score of 0 for edge to rightmost child and keep increasing by 1 towards left. Score of each word = sum of edge scores from the root to that word.
   Overall Score = sum or max or mean
2. Frazier Scoring : trace a path up the tree from a word upto either the root or a node that's not the leftmost child of its parent. Overall Score = sum or max or mean

1.1.4 Features based on Sentiment Analysis

Tag every tweet as one of {positive, negative, neutral} sentiments using Stanford CoreNLP [2].

1.1.5 Features based on N-grams and word counts

1. Length of tweets, Length of words, number of hashtags
2. N-gram features: N most important ngrams based on their TF-IDF scores.
3. Linguistic Inquiry and Word Count (LIWC): Previous studies using LIWC have found that it is possible to characterize depression through natural language use [3]. We plan to use word counts for each LIWC [4] category as features in our supervised classifier.

1.2 Topic modeling

Topic models are effective unsupervised tools to study hidden structure of texts. By visualizing word clusters detected by topic models, we may be able to characterize distinctions between language of the normal group and that of the schizophrenic group. However, for topic models to be effective on tweets, which are mostly brief, one important preprocessing step is to group tweets within a period of time to obtain longer documents.

After having those documents, for the first step, we want to run a basic latent dirichlet allocation model (LDA) on them. We compare the difference between the top words in topics identified from documents of normal group and those of the schizophrenic group. Also, using the topic distributions of documents, we can compute on average how often each topic is mentioned. The hope is to discover idiosyncratic topics of schizophrenic people.

In the second step, following the intuition that mental illness is closely related to sentiments, we conduct similar experiments on topic models that incorporate sentiments into their structures. The sentiments can be approximated by a Twitter sentiment detector ([5]). We are considering three possible sentimental topic models:

1. An LDA-like model that splits each topic into two versions: one positive and one negative. This model allows studying how different the topics are mentioned in a positive versus a negative context.
2. Also an LDA-like model but instead the words are split into two versions: one positive and one negative. In other words, we attach the sentiment of a sentence to all of its words. For instance, using this model, we can study what words co-occur with "family" when it is mentioned in a positive context versus negative context.
3. Use a supervised LDA model ([6]) with sentiment as the reponse variable. In the end, this model assigns a sentiment score to each topic. It would be interesting to juxtapose scores of overlapped topics between the two groups (if there are any).

2 Classification

For classification the task is, given all the tweets of a user, predict if he is at risk. We can think of the following general strategies:

1. Use per user features to train classifiers
2. Average per tweet features over a user and then train classifiers.
3. Explore ways to fuse features that are somewhat different in nature (For example, concatenation)

Specifically we plan to use the following:

1. Build an initial n-gram based model.
2. Add above mentioned features and compare the model's performance using F1-score/ AUC.
3. Build an ensemble model using different classifiers like Logistic Regression, SVM, Random Forest 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

[2] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010

[3] N. Ramirez-Esparza, C. K. Chung, E. Kacewicz, and J. W. Pennebaker, "The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches." in *ICWSM*, 2008.

[4] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," 2010. [Online]. Available: http://homepage.psy.utexas.edu/homepage/students/Tausczik/Yla/index.html

[5] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, vol. 1, p. 12, 2009.

[6] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *Advances in neural information processing systems*, 2008, pp. 121–128.

We have read and understood the conditions on proper use of the Qntfy dataset.