**Suicide Watch Project**

Colin King, Prudhvi Gurram, Pouya Moetakef

1. Introduction

The predicted roles for each group member is listed below:

Colin King:

Prudhvi Gurram:

Pouya Moetakef will work on topic models and word classification explorations including subjectivity and emotions of the words.

For this project we have decided to:

1. Explore word classification, since individual words carry meanings that reveal the state of mind of the writer.
2. Explore topic modeling to understand differences between posts (positive or negative) and be able to further filter unrelated posts from the positives, for higher performance of the supervised classifier.
3. Data and Methods:
   1. Data sources and usage

We are planning to use the reddit data as well as the personality test data. To use positive and controls posts data, the sentences need to be tokenized. Looking at the data there are good amount and emoticons used. We are planning to use two approaches: (i) keep emoticons unchanged and study how they show up as differences between positive and control data set. (ii) Replace them with their respective feeling words. For instance :) will replace with “happy” or :( will replace with “sad”. This way, emoticons can be used in topic models and emotion classifications.

As suggested by guidelines, the data set will be separated to train, dev, and test using the user ids given in txt files.

To work with LIWC lexicons, regular expression will be used to match the given words (containing \*) with respective words read from the data.

* 1. Planned Methods for exploratory data analysis

Baseline:

For determining the Baseline for this project we will use bigram classification. The positive posts are given class +1 while negative posts will be given class -1. Then through simple bigram analysis a baseline for the predication of positive class can be determined. The performance of the baseline can be improved by incorporation of interpolation for 3-gram classification, where weighted monogram, bigram, and 3-gram are used and the weights are adjusted such that to maximize the probability of dev-set. Comparing highest frequency bigrams in each class with those reported in other papers can be a start for determining features.

Word Classes:

As stated by (Milne et. al. 2016), subjectivity and polarity of a post can be used as a feature in determining the state of the mind. The hypothesis says that negative polarity is being used more in the positive posts than in controls. So, using the clues given in this lexicon may help us distinguish the differences better and also can be used as a feature in supervised classification. Therefore, we plan to use the MPQA (Multi Perspective Question Answering) package provided via <http://mpqa.cs.pitt.edu/> website for subjectivity and polarity test of users’ posts.

An exploratory feature that was suggested by (Mowery et. al. 2017 and De Choudhary et. al. 2016) is the focus on self (more use of “I”) in the context. We intend to explore this observation for the reddit posts as well, and compare the focus on 1st, 2nd, and 3rd person between positive and control posts. Therefore, statistical significance will be determined using t-test or Wilcoxon signed rank test using SciPy package.

As mentioned by (Mowery et. al. 2017), the Pennebaker’s Linguistic Inquiry and Word Count (LIWC) lexicons associated with negative emotions can be used to assess the difference between positive and control posts. Here a hypothesis is that the positive posts will have higher association with negative emotions. Honestly, this hypothesis is highly expected.

Topic Models:

Another exploration we have planned is to use the LDA topic modeling tool (as we have practiced before using MALLET) to identify the topics that can best represent the posts. The top contributing words in each topic will be used to determine the word class using LIWC for proper class label assignment. The number of classes will be tweaked to minimize the overlapping of the classes. However, since this study is performed to determine rough class assignment, the number of topics will be tweaked between 5 and 20, with steps of 5. We expect that this analysis to give us clues regarding filtering posts that are nor associated with negative emotion topics.

The similarity topic between positive corpus and negative corpus can be another assessment for determining differences between the two. The averaged DL divergence will be used to assess this similarity. For this test, the 100 posts from the positive and 100 posts from the control will be chosen randomly to generate a heat map of similarity. It is expected that positive posts show higher similarity between themselves compared with controls. Then top 50% performing posts will be kept as reference (this process can be refined by repeating multiple times, i.e. replacing the 50% lowest performing posts with randomly selected posts, to reach a more uniform similarity). The top contributing words in this test can be identified as feature selection. Furthermore, this test can be used to further filter the positive posts. The posts that are least similar to the reference can be tagged as false and removed from supervised training for better performance.

Readability measures:

As mentioned in the guidelines and also (De Choudhary et. al. 2016) readability of user post can show degree of mental health. Therefore, we plan to explore this option using the package source provided at <https://pypi.python.org/pypi/readability/0.1>. The features we would like to consider in this test are readability grades such as SMOG Index for the complexity of the sentence, and sentence info measure such as characters per word, percentage of complex words, or words per sentence, where information regarding use of short sentences or long sentences can be beneficial to assess the differences.

* 1. Planned methods for supervised classification

For supervised classification we plan to use the linear regression classifier or perceptron to predict the appropriate class. Since, all the posts from a user are not usually relevant, a filtering process as described before using similarity measure between posts will be used. Further filtration will use of a time window where the post regarding suicide is observed. It is likely that posts that are more than a week away from the target post might not be relevant. Once the message filtration is done, the remaining posts will be grouped together based on the topic classes identified through LDA topic modeling study. The model is then trained on the training data set for multiclass classification.

To weight the topic classes properly, we intend to use the data provided through mypersonality\_test dataset. Since, the scores of depression and neuroticism can be trained on the keywords from each topic class, we end up with rough values for weights for multi-class classifier. The validity of these weights will be assessed through dev-data set. If poor performance was observed, a retrain on the training set will be use to tweak the weights.

As suggested in the guidelines, classification will be evaluated using precision, recall, and F-measure. We also, would like to try a 10-fold cross evaluation over the training set and dev set, to assess if the held-out set (dev set) contains posts that may improve the our supervised classifier.

Reference:

[De Choudhury et al., 2016] De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., and Kumar, M. (2016). Discovering shifts to suicidal ideation from mental health content in social media. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pages 2098{2110. ACM.

[Milne et al., 2016] Milne, D. N., Pink, G., Hachey, B., and Calvo, R. A. (2016). Clpsych 2016 shared task: Triaging content in online peer-support forums. In Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology, pages 118{127, San Diego, CA, USA. Association for Computational Linguistics.

[Mowery et al., 2017] Mowery, D., Smith, H., Cheney, T., Stoddard, G., Coppersmith, G., Bryan, C., and Conway, M. (2017). Understanding depressive symptoms and psychosocial stressors on twitter: A corpusbased study. Journal of Medical Internet Research, 19(2).