Suicide Watch Project

Colin King, Prudhvi Gurram, Pouya Moetakef

1. Introduction

In this project, we attempt to answer the question of whether we can predict or assign a likelihood to a user becoming suicidal based on linguistic signal in their social media posts We present various measures and methods by which we identify differences between the negative posts and the positive posts. The intention is to observe the trend of certain characteristics in a user's posts and predict if they are going to post in SuicideWatch (SW).

The predicted roles for each group member is listed below:

Colin King will establish a baseline with a simple word-level n-gram language model. He will also explore character n-gram language models, since this may handle intentional misspellings (ex. "saddd") and emoticons. Colin will also explore for bigram collocations that appear at significantly different frequencies between the positive and control posts. Finally, Colin will explore the possibility of categorizing each subreddit, in hopes of identifying posting trends by positive users leading up to their eventual post in SW.

Prudhvi Gurram will work on readability index, time-bucketing and pronoun analysis.

Pouya Moetakef will explore word classification, including subjectivity and emotions of the words, since individual words carry meanings that reveal the state of mind of the writer. Furthermore, he will 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.

2. Data and Methods:

2.1. Data sources and usage

We are planning to use the Reddit data dump as well as the MyPersonality data. Most of our preprocessing will consist of simply tokenizing the input, however for our topic model and word class analysis', we will replace emoticons with their associated descriptive word. For example, the emoticon ":)" would be replaced with "happy." All other features benefit do not require this preprocessing step, or benefit from the added semantic meaning that emoticons provide.

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

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

All training for this project will be conducted solely on user posts prior to each user's first post on SW, since once a user posts on SW, it is likely that they have already made an attempt at suicide.

During training, we will also exclude any posts to the following set of mental health subreddits: Anger, BPD, EatingDisorders, MMFB, StopSelfHarm, addiction, alcoholism, depression, feelgood, getting over it, hardshipmates, mentalhealth, psychoticreddit, ptsd, rapecounseling, socialanxiety, survivor-sofabuse, and traumatoolbox

2.2. Planned Methods for exploratory data analysis

Baseline:

For the baseline for this project, we will use a word-level n-gram language model. Specifically, we will use interpolation with unigram, bigram and trigram models, where the respective weights are identified by maximizing the probability of the dev-set. A separate language model will be constructed for each of the positive and negative datasets. During testing, a label will be assigned to each set of test posts, depending on which language model has the higher probability of producing those posts. We intend to identify which n-grams have the highest correlation with the positive and negative classes, in hopes that this may identify new potential 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 used more in the positive posts than in the control posts. So, using the clues given in this lexicon may help us distinguish the differences and serve as a feature in our supervised classification model. Therefore, we plan to use the MPQA (Multi Perspective Question Answering) package provided via http://mpqa.cs.pitt.edu/ website for subjectivity and polarity analysis of users' posts.

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.

Character N-gram Language Model:

As mentioned in (Coppersmith et al. 2015), character n-gram language models provide a handful of benefits over the traditional n-gram language models that we are using in our baseline. Specifically, they can handle the creative language use and emoticons that is common to social media, along with providing some robustness to misspelling. We will test character models that use different sized sliding windows, from 3 characters to 6 characters, and choose the window sizes that performs best on the dev-set. Similar to Coppersmith et al, we will normalize the dataset by lowercasing all characters, tokenizing usernames with an "@" symbol, and tokenizing URLs with a "*" symbol. We hope to compare the success of a character n-gram model with that of the baseline word-level n-gram language model to identify if the improvement identified in the Twitter dataset of (Coppersmith et al. 2015) can be replicated in the Reddit dataset.

Pronoun Analysis:

An exploratory feature that was suggested by (Mowery et. al. 2017 and De Choudhary et. al. 2016) is the focus on self (increased use of "I"). We intend to explore this observation within the Reddit posts as well, and compare the focus on 1^{st}, 2^{nd}, and 3^{rd} person pronouns between positive and control posts. The usage/focus on different pronouns in the posts might signify the difference between positive and control posts. To quantify the difference in usage between positive and control posts, we will test the statistical significance using t-tests or Wilcoxon signed rank tests using SciPy.

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 positive and control 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 word overlap. 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 hope that this analysis will give us clues that could lead to methods to filter out posts that are not associated with negative emotion topics.

The similarity of topics between the positive corpus and negative corpus can be another assessment for determining differences between the two. The averaged KL-divergence will be used to assess this similarity. For this test, 100 posts from the positive corpus and 100 posts from the control corpus will be randomly chosen to generate a heat map of similarity. It is expected that positive posts will 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.

Subreddit Classification:

We would like to perform a linguistic analysis of the descriptions of each subreddit to attempt to place it in a predefined (to be determined) set of classes. We hope to be able to identify changes over time, such that if a user stops posting in r/movies (possibly tagged as "entertainment") and instead increases posting in r/TIFU (possibly tagged as "negative-stories"), then this may indicate a decrease in happiness or life outlook.

Collocation Analysis:

We hope to explore whether users who eventually post in SW use certain phrases more often than control users. To test this, we intend to analyze Reddit posts to identify bigram collocations. We will perform a statistical significance test to verify any claims we make.

Readability measures:

As mentioned in the project specification, and also (De Choudhary et. al. 2016), readability of user posts can indicate the degree of mental health. The features we would like to consider in this test are readability grades such as the SMOG Index the measure sentence complexity, and statistics such as characters per word, percentage of complex words, or words per sentence. We will use the Python readability package (https://pypi.python.org/pypi/readability/0.1).

Time-bucketing:

We hypothesize that a trend in the time of posting may be indicative of lifestyle changes which may indicate mental issues. On its own, an increase in late-night postings is unlikely to be indicative of suicidal thoughts, but in combination with other features, it is possible that it could increase our model accuracy. Since we do not have access to user time zones, we will bucket times into "morning", "afternoon", "evening", "late-night" and "early-morning" based on the US central time zone. Our hope is that this bucketing approach will capture a majority of late-night postings, with enough leeway to handle users in other time zones.

Interaction variables:

Interaction variables, mentioned in (De Choudhary et. al. 2016), include further measurement categories which include, volume of posts, post length, volume of comments, and mean vote difference. This might give us a picture of how interactive and responsive the user is. We intend to test these variables as features in our supervised classification model to see if it aids with distinguishing positive users from control users.

2.3. Planned methods for supervised classification

Based on our feature exploration above, we will retain the set of features that identify statistically significant differences between the positive and control corpuses. Each feature must produce a probability likelihood, which will be combined in a weighted linear classifier which will produce an overall likelihood of a user eventually posting in SW. The weights for this model will be trained with K-fold cross validation on the combined training and dev sets, in case there exist documents in the dev-set that can improve the

quality of the model. Features that analyze trends over time will have access to all posts up to and excluding the post on SW, however all other features will receive a discourse consisting of all posts in the week leading up to a post on SW concatenated together. This is done since we hypothesize that most linguistic signal will be expressed very close to the actual attempt, rather than over many years.

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

If time permits and all the above were accomplished, then we intend to use neural networks, where we input the vector representation of the control posts for a given user and predict if the user is likely to post on SW.

References:

[Coppersmith et al. 2015] Coppersmith, G., Leary, R., Whyne, E., and Wood, T. Quantifying suicidal ideation via language usage on social media.

[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).