CMSC 773 Project Proposal

Jo Shoemaker and Julian Vanecek

April 12, 2018

1 Introduction

This group is comprised of Julian Vanecek and Jo Shoemaker. Given the small size of this group, we plan to contribute roughly equally to all the tasks described below.
We first hope to find the most predictive features during our data exploration phase, and then to build a logistic regression model to classify users as either positive or negative instances of "high risk for suicide within two weeks." We chose this two week cutoff for two reasons: first, suicidal crises on average last for 1-2 weeks (Witte et al., 2005), and so it makes more sense to classify a person based on this timeframe, rather than once across their entire history. Second, by considering snapshots of user behavior in two-week windows rather than all at once, we can increase the number of training instances to the model (i.e. we go from having one training instance per user n to having $n \times \frac{w_n}{2}$ instances, where w_n is the number of weeks that user n has used `reddit`. For labeling purposes in our training, a positively CrowdFlower-labelled user's behavior-snapshots will not be labelled as a positive instances of suicide risk until two weeks before their first post in `r/SuicideWatch`, and won't be labelled as such two weeks after their last post to this subreddit.

2 Data and Methods

2.1 Exploratory Analysis

Using Pearson's r, we would like to test the strength and significance of correlation between suicide risk (according to CrowdFlower ratings on 4-point scale) and the following potentially interesting features:

- **Ratio of singular personal pronouns to all other pronouns.** This has widely been considered an approximation for inward-focus, which is higher among the depressed and suicidal (Vioulès et al., 2018). This measure can easily be retrieved using regular expression matching.
- **Ratio of present tense verbs to all verb tenses.** We have a hunch that suicidal people may be more focused on the present than on the past or future. We will extract this feature using a POS tagger (probably Stanford's) that categorizes verbs by tense.
- **Percentage of vocabulary items from the following LIWC categories:**
  - "Anger"
  - "Sad"
  - "Health"
  - "Sexual"

  - "Money"
  - "Death"
  - "Friends"
  - "Family"

- **Degree of "linguistic accommodation".** This feature was determined to be highly correlated with suicidality by De Choudhury et al. (2016). We will use their reported strategy for measuring linguistic accommodation.

- **Readability.** This feature was determined to be highly correlated with suicidality by De Choudhury et al. (2016). We will use their reported strategies for measuring readability.

- **'Mental health thematicity' of subreddits the user contributes to.** We suspect that certain subreddits, while not explicitly about mental health, will allude to the same themes. Using word2vec vectors, we will create an average vector representation of mental health subreddits. We will measure cosine similarity of this vector to an averaged vector of non-explicitly mental-health-related subreddits which the user contributes to.

- **Thematic fit of contributions to subreddits.** A suicidal user might be more likely to post erratically or without social awareness. Given a word2vec average vector representation of a subreddit, we will report the cosine similarity of this vector with a vector representing an individual user's posts on that subreddit.

- **Spelling accuracy.** A high degree of misspelled words might indicate distractedness or impulsiveness. We will run a user's input through a spell checker to figure out their percentage of misspelled words.

We also want to investigate the following non-linguistic measures:

- **Time distribution of posts.** We suspect that posting throughout the day might indicate low community engagement and isolation, which is correlated with depression. We will operationalize this as high average standard deviation in times of posts for the same days of the week.

- **Frequency of posting.** De Choudhury et al. (2016) found that suicidal individuals only posted about half as frequently as non-suicidal individuals.[1]

- **Average post length.**

- **Standard deviation of post length.**

Finally, suicidality (at least for first-time attempters (Witte et al., 2005)) is usually brought on by sudden changes, rather than by ongoing but stable bad situations. Because of this, for all the features x listed above we will consider the "change in x" in a user's behavior in the last two weeks compared to their aggregated previous behavior.

[1] They were, however, just as likely to comment.

2.2 Final Model

For whichever of the above features turn out to be most predictive, we would like to train a neural logistic regression model to binarily classify training instances. Depending on if time allows, we might try training models with increasing complexity, including convolutional layers similar to Yates et al. (2017). Since we are extracting features manually, we would not expect to need as many layers as in that paper.

3 Evaluation

We will evaluate intermediate models trained on the training set using the dev set in order to tune our hyperparameters. Additionally, we would like to use an objective function that penalizes low recall more heavily than low precision, but we haven't decided on one yet.

References

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.

Vioulès, M. J., Moulahi, B., Azé, J., and Bringay, S. (2018). Detection of suicide-related posts in twitter data streams. *IBM Journal of Research and Development*, 62(1):7–1.

Witte, T. K., Fitzpatrick, K. K., Joiner, T. E., and Schmidt, N. B. (2005). Variability in suicidal ideation: a better predictor of suicide attempts than intensity or duration of ideation? *Journal of Affective Disorders*, 88(2):131–136.

Yates, A., Cohan, A., and Goharian, N. (2017). Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.