

Lab6_Presentation

Linda Gai

10/14/2016

Introduction

I am working on a project on mental health (specifically suicidal ideation) and trolling among Reddit users, along with Lacey. Generally, we are interested in whether we can identify users that are depressed, and their identifying characteristics. Our ultimate goal is to come up with models of behavior for users that admit to having suicidal ideation on Reddit's r/depression subreddit. That is, we plan to look into the suicidal users' post history to answer questions about their behavior. Specifically, I am interested in developing a logistic regression model to predict whether a user will make a post containing thoughts of suicidal ideation, based on their post history. To do this, I will first examine the word choice and sentiment differences between suicidal and non-suicidal users.

Reddit is a popular online message board, commonly used to share news and images, as well as discussion on user-created topics. It is the 11th most popular website in the US, ranked just under Twitter which comes in at #10. Notably, some discussion boards (known as "subreddits") are used to provide and ask for advice and support, and many users share deeply personal information and emotional states. We are particularly interested in how we might be able to identify suicidal ideation among Reddit users, as many users post suicide notes and/or reveal intention to commit suicide.

Methods

Collected data set(s)

We collected the following datasets:

- all the comments from r/depression from June 1-30, 2016, in JSON format
- a large training set of approximately 600 comments from mental health subreddits, including but not limited to r/depression. The comments were obtained and scored as follows:
 - ~110 submissions (i.e. comments from the original poster), scored by both Lacey and Linda, randomly sampled from a set of ~20 of the largest mental health subreddits excluding r/depression, using the PRAW python package's random sampling function. Comments were scored for current suicidal ideation, past suicidal ideation, and the presence/absence of mood or anxiety disorders. ~50% of these posts indicate suicidal ideation.
 - ~500 comments (consisting of both submissions and replies) from r/depression that were posted in June. Lacey and Linda each scored ~300 comments, with ~100 comments scored by both. Comments were scored only for the presence of suicidal ideation, and were obtained from June posts in r/depression. Only about 30 of these comments indicate suicide ideation.
 - a JSON parser (written with help from John) that takes in a JSON file and spits out a data frame that contains the attributes of a submission, along with all of its replies

Analysis

Classification of Posts into Suicidal and Non-suicidal Categories

We created a classification algorithm that will search individual posts for phrases that match phrases that indicate suicidal ideation. The idea is that, since we are only interested in the behavior of suicidal users, we want an algorithm that has very high sensitivity (i.e. will identify only true positives of suicidal users), but we don't care too much about the specificity of the algorithm.

Exploratory Data Analysis

How does word choice differ between suicidal and non-suicidal users?

Here, I compare word tokens between suicidal and non-suicidal POSTS (not comments), using the classified set that has already been scored by the algorithm.

```
## Loading required package: NLP

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##   date

##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:NLP':
##
##   annotate

## Joining, by = "word"

## # A tibble: 6 × 4
##   word      sui    not.sui    diff
##   <chr>    <dbl>    <dbl>    <dbl>
## 1 feel 0.016862409 0.022670404 -0.005807995
## 2 depression 0.007011212 0.010956906 -0.003945694
## 3 time 0.011212023 0.013869328 -0.002657306
## 4 people 0.010117445 0.012447301 -0.002329855
## 5 feeling 0.003313315 0.005359951 -0.002046635
## 6 day 0.006715380 0.008226795 -0.001511415
```

```
## # A tibble: 6 × 4
##   word      sui    not.sui    diff
##   <chr>    <dbl>    <dbl>    <dbl>
## 1  dont 0.003549981 0.002055559 0.001494421
## 2 fucking 0.005384138 0.003860441 0.001523697
## 3 anymore 0.005916635 0.004006290 0.001910345
## 4  live 0.005709552 0.003081060 0.002628492
## 5  kill 0.005472887 0.001267063 0.004205824
## 6   die 0.006389965 0.001449374 0.004940591

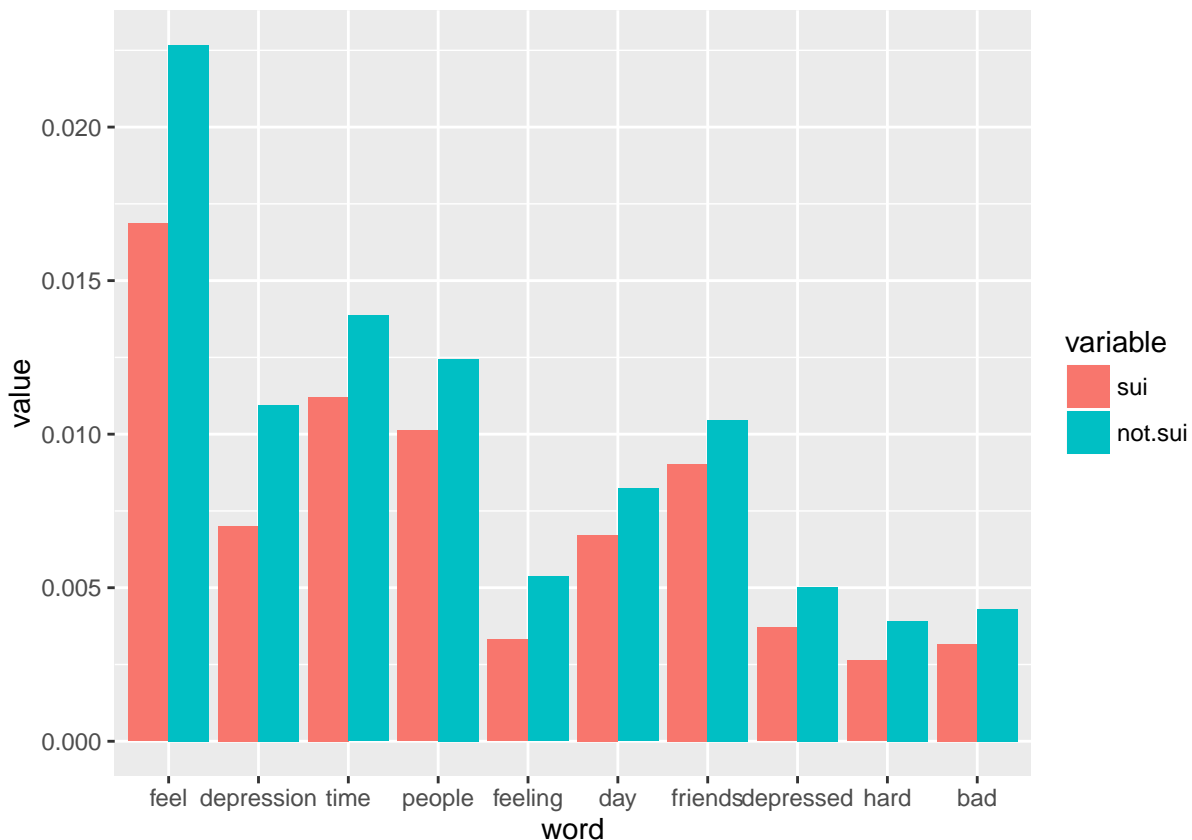
## [1] -0.005807995  0.004940591

##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##   smiths

## Using word as id variables

## [1] feel      depression time      people    feeling    day
## [7] friends   depressed hard      bad      feel      depression
## [13] time      people    feeling    day      friends    depressed
## [19] hard      bad
## 10 Levels: feel depression time people feeling day friendsdepressed hard bad
```



What are the overall sentiment pattern differences between suicidal and non-suicidal posts?

```
## Joining, by = "word"
## Joining, by = "word"

## -----

## data.table + dplyr code now lives in dtplyr.
## Please library(dtplyr)!

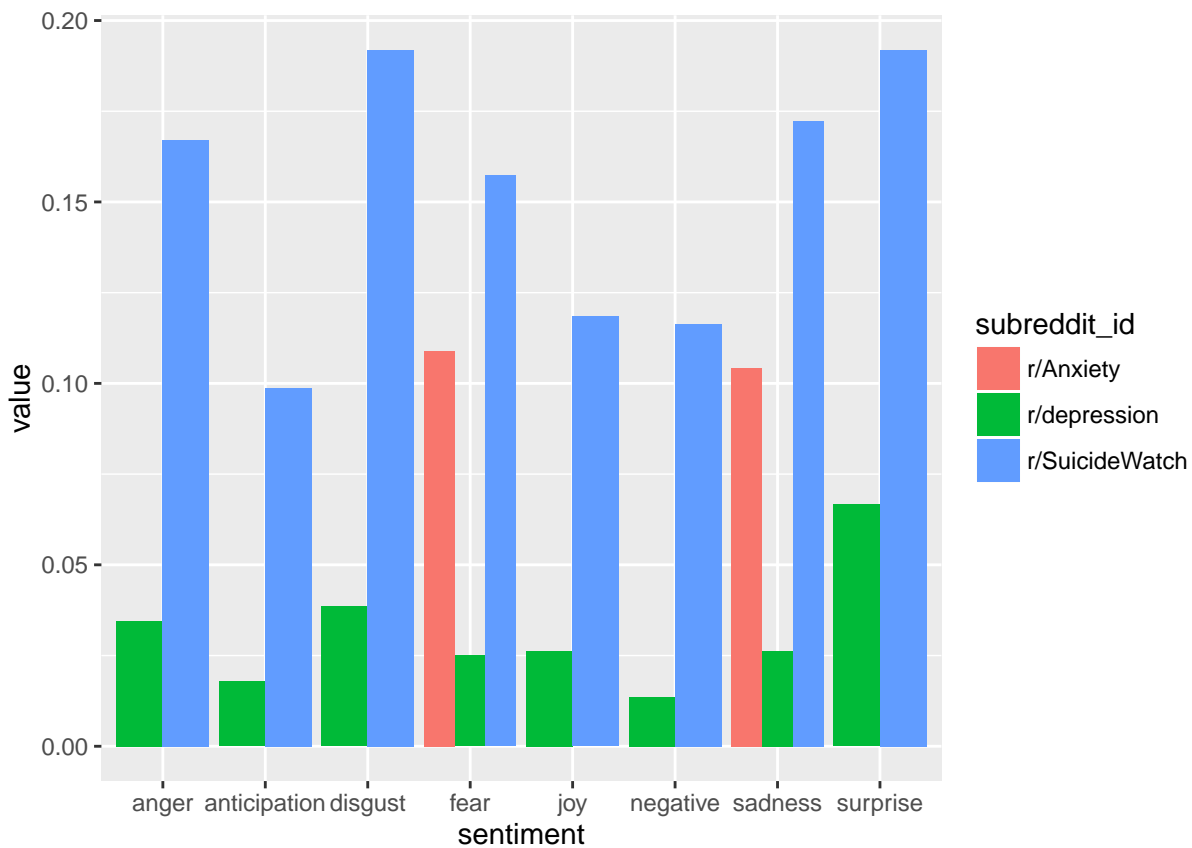
## -----

##
## Attaching package: 'data.table'

## The following objects are masked from 'package:reshape2':
##
##   dcast, melt

## The following objects are masked from 'package:lubridate':
##
##   hour, mday, month, quarter, wday, week, yday, year

## The following objects are masked from 'package:dplyr':
##
##   between, last
```



What are the other subreddits that suicidal users post in?

```
load(file="Classified_Posts/June_Authors_Data.RData")

not.dep.sent <- post.sent %>%
  filter(subreddit_id!="t5_2qqqf") %>%
  group_by(sentiment,subreddit_id) %>%
  summarize(n=n()) %>%
  filter(n>100) %>%
  arrange(desc(n))

x <- sort(table(user.post.data$subreddit_id),decreasing=T)[1:10]
names(x)<-c("depression", 'AskReddit', 'datfeel', 'SuicideWatch',
           'Fireteams', 'asktransgender', 'leagueoflegends', 'DotA2',
           'Anxiety', 'ForeverAlone')
x
```

##	depression	AskReddit	datfeel	SuicideWatch
##	2596	817	339	294
##	Fireteams	asktransgender	leagueoflegends	DotA2
##	234	212	207	179
##	Anxiety	ForeverAlone		
##	172	170		

What are the sentiment pattern differences in posting history between suicidal and non-suicidal users?

Logistic regression model for predicting suicidal ideation on Reddit