Computational Linguistics II Final Project:
Exploring Linguistic Signal for Suicidality in Social Media

An important note

In this project, you will be looking at posts written by users in an online discussion forum. The dataset comes from a project where the goal is finding new ways to help prevent suicides.

Before you go any further, please recognize that some of the posts you might see in this project were written by people in real distress, and they can be difficult or upsetting to read. **If you think that you might be affected personally in a negative way by doing this project, please err on the side of caution and stop here; do *not* do the project — an alternative project will be provided. If you start the project and you find that it's upsetting, similarly, please stop and contact the instructor to make an alternative project plan.**

If you're feeling like you (or someone you know) could use some support or assistance, please take advantage of one of the following resources:

- National Suicide Prevention Lifeline: 1-800-273-8255 (TALK).
  - Veterans please press 1 to reach specialized support.
- Spanish: 1-800-SUICIDA
- Crisis Text Line: Text "START" to 741-741
- Online chat: http://www.suicidepreventionlifeline.org/gethelp/lifelinechat.aspx
- https://www.reddit.com/r/SuicideWatch/wiki/hotlines - This page provides information about phone and chat hotlines and online resources in the U.S. and worldwide.

Please note that all the posts you will encounter in this project were anonymous — not even the researchers who created the dataset know who these people are, and the posts were made over a period of years. Although it's tragic that there is no direct way for us to help the people who have written these posts who may be at risk of suicide, research progress on this dataset is aimed at better understanding the factors connected with suicide attempts, using that information to do a better job assessing risk, and hopefully contributing to more effective ways of getting people help.

Also, although the posts we're working with are anonymous, it is absolutely essential that you read and understand Section 6, below, on the ethics of working with social media data.

1 Introduction

In the United States, mental health problems are among the costliest challenges we face, and the World Health Organization (WHO) has reported that mental illnesses are the leading cause of disability adjusted life years worldwide. The numbers are staggering: to cite just a few, between 1996 and 2006, annual expenditures on mental disorders in the U.S. rose from $35.2B to $113B,[1] some 25 million American adults will have an episode of major depression this year,[2] and suicide is the third leading cause of death for people between 10 and 24 years old.[3] Alzheimer's disease is a deep concern for an aging baby-boomer generation, and, at the other end of the age spectrum, we are just beginning to recognize the effects of mild traumatic brain injury on young athletes. Diagnosis of autistic spectrum disorders has risen to include 1 in every 88 American children.[4] Depression is estimated to account for a third of worldwide disability costs, and schizophrenia ranks higher in costs than congestive heart failure and stroke (Insel 2008, Soni 2009).

The importance of mental health as a problem space cannot be overstated.

For clinical psychologists, language plays a central role in diagnosis and in monitoring of patients. Indeed, many clinical instruments fundamentally rely on what is, in effect, manual coding of patient language. For example, in assessment for formal thought disorders, analysis of natural speech is an essential factor in the diagnosis, as the clinician must assess the patient's language for diagnostic features such as incoherence, derailment, loose associations, and tangentiality (Association, 2013). Applying language technology in this domain, e.g. in language-based assessment, could potentially have an enormous impact, because many individuals are motivated to underreport psychiatric symptoms (consider active duty soldiers, for example) or lack the self-awareness to report accurately (consider individuals involved in substance abuse who do not recognize their own addiction), and also because many people — e.g. those without adequate insurance or in rural areas — cannot even obtain access to a clinician who is *qualified* to perform a psychological evaluation (APA, 2013; Sibelius, 2013). Bringing language technology to bear on these problems could potentially lead to inexpensive screening or monitoring methods that could be administered by a wider array of healthcare professionals, which is particularly important since the majority of individuals who present with symptoms of mental health problems do so in a primary care physician's office. Given the burden on primary care physicians to diagnose mental health disorders in very little time, the American Academy of Family Physicians has recognized the need for diagnostic tools for physicians that are "suited to the realities of their practice".[5]

This project focuses on suicidality. The majority of assessment for suicide risk takes place via in-person interactions with clinicians, using ratings scales and structured clinical interviews (Batterham et al., 2015; Joiner Jr et al., 1999, 2005). However, such interactions can take place only after patient-clinician contact has been made, and only when access to a clinician is available. This is no small challenge in many places — in the U.S., for example, nearly 124 million people live in federally designated mental health provider shortage areas, where access to a provider can be difficult even when the person (or someone close to them) knows that clinical help is needed (Bureau of Health Workforce, 2017).

An emerging subset of the artificial intelligence and language technology communities has been making progress on automated methods that analyze online postings to flag mental health conditions, with the goal of being able to screen or monitor for suicide risk and other conditions (Calvo et al., 2017; Resnik et al., 2014; Milne et al., 2016; Milne, 2017). This dovetails with the fact that people are spending an increasing amount of their time online, and in fact online discussions related to mental health are providing new opportunities for people dealing with mental health issues to find support and a sense of connection; these include Koko,

[1] Yes, that's 'B' for billion: `http://www.nimh.nih.gov/statistics/4TOT_MC9606.shtml`,`http://www.washingtonpost.com/blogs/wonkblog/wp/2012/12/17/seven-facts-about-americas-mental-health-care-system/`

[2] `http://www.nami.org/Template.cfm?Section=depression`

[3] `http://www.cdc.gov/violenceprevention/pub/youth_suicide.html`

[4] `http://www.cdc.gov/ncbddd/autism/data.html`

[5] http://www.aafp.org/afp/1998/1015/p1347.html

itskoko.com; ReachOut, reachout.com ; 7cups, 7cups.com; Reddit, reddit.com, and others. Although many such discussions are peer-to-peer, site moderators often play a crucial role, identifying users who may be at imminent risk and require intervention. Technological tools for analyzing peoples' online postings (subject to ethical and privacy issues, discussed below) have enormous potential both for *screening* (a binary decision that someone should be evaluated in terms of suicide risk) and for *risk assessment* (evaluating someone in order to assign a level of risk).

This final project focuses on screening people for suicidality based on social media postings. It is intended to provide you with the opportunity to exercise what you have learned in class on a challenging, open research problem. In this document, we'll describe the data, along with the basic goals of the project. (And, of course, how you'll be graded.) *There are no guarantees that you will get sensible or interpretable results.* But that's ok: what matters is how thoughtfully you approach things, how much you demonstrate mastery of ideas and techniques that we've learned about over the course of the semester, and how carefully and coherently you describe what you did.

2 Background on the problem

2.1 Prior work to draw on

There are too many references in here for you review all of them, but I'm erring on the side of too much rather than too little information. I'll try to provide some guidance as to the most useful things to look and I'm happy to answer questions; please feel free to also use the class discussion board to talk about this.

Calvo et al. (2017); Guntuku et al. (2017) present reviews of NLP research in which social media are used to identify people with psychological issues who may require intervention, and Conway and OConnor (2016) provide a shorter survey focused on public health monitoring and ethical issues, highlighting the annual Workshop on Computational Linguistics and Clinical Psychology (CLPsych), initiated in 2014, as a forum for bridging the gap between computer science researchers and mental health clinicians (Resnik et al., 2014). Recent CLPsych shared tasks using data from the ReachOut peer support forums have provided opportunities for exploration of technological approaches to risk assessment and crisis detection (Milne et al., 2016; Milne, 2017).

This project is specifically about screening for suicidality based on a person's social media postings. Although predictive modeling in mental health is a burgeoning area, a key challenge for work on mental health in social media is connecting the clinical side with available social media datasets. Combining ground truth health record data with social media data is rare, with Padrez et al. (2015) representing a promising exception; they found that nearly 40% of 5,256 Facebook and/or Twitter users who were approached in a hospital emergency room consented to share both their health record and social media data for research.[6] Approximations of clinical truth are more common, e.g. self-report of diagnoses in social media (Coppersmith et al., 2014), or observed user behaviors such as posting on SuicideWatch (De Choudhury et al., 2016). Coppersmith et al. (2015b, 2016) employed a Twitter data collection method (Coppersmith et al., 2014) discussed below to discover Twitter users with self-stated reports of a previous suicide attempt in order to identify valuable signal and support automated classification.

Two pieces of recent work seem particularly relevant and worth looking at. Yates et al. (2017) won a best paper award at the 2017 EMNLP conference (one of the top-tier conferences in NLP) for their work on depression and self-harm risk assessment in online forums. This paper is a great example to follow of high quality work in this domain that obtained strong results. Even more recently, Vioulès et al. (2018) applied a similar data collection approach to the one we've taken here, in their case searching Twitter for

[6]Interestingly, participants agreeing to social media access were only slightly younger on average than those who declined (29.1 ± 9.8 versus 31.9 ± 10.4 years old.

tweets containing key phrases based on risk factors and warning signs identified by the American Psychiatric Association and the American Association of Suicidology. They report results for human annotation as well as predictive modeling.

2.2 Risk factors for suicidality

Identifying risk of suicide accurately takes a great deal of training, and the factors contributing to suicidality are not fully understood.[7] However, clearly depression is a major factor. There has been some significant recent work exploring potential indicators of depression in social media. In particular, see Mowery et al. (2017) for a detailed corpus study looking at, for example, the predictive value of depression-related keywords, and at correlations between depressive symptoms and psychosocial stressors. They find, as an example, that fatigue or loss of energy (symptom) correlates with disturbed sleep and educational problems (stressor). The Mowery et al. study references the 2015 CLPsych shared task, which involved screening for depression in social media users (Coppersmith et al., 2015a).

Although depression is a major factor, many more people experience depression than actually attempt suicide. From a clinical perspective, there are a number of additional factors that people with clinical training take into account when judging risk. Some of these may not be obvious — for example, research shows that someone who is showing signs of agitation can be at higher risk than someone who just seems down or depressed (Popovic et al., 2015). Informed by discussion with several experts in assessment of suicidality, we group a number of relevant factors that contribute to higher assessments of suicide risk into thoughts, feelings, logistics, and context, expanding on work by Corbitt-Hall et al. (2016) that provided lay definitions based on risk categories in Joiner Jr et al. (1999). (See also Vioulès et al. (2018) for their discussion of risk factors and warning signs.)

- Thoughts
  - Thinking about suicide, having suicide on their mind
  - Having told friends or family they are thinking about suicide
  - Feeling that they are a burden to others
  - Endorsement of suicidal beliefs, even without the word suicide (e.g., I deserve to die, I can never be forgiven for the mistakes I made)
  - A "fuck it" (screw it, game over, farewell) thought pattern
- Feelings
  - Expressing lack of hope for things to get better
  - A sense of agitation, not being able to stand still physically or mentally
  - Indications of being impulsive; risky behavior (e.g. reckless driving, promiscuity)
- Logistics
  - Talking about plans that involve suicide
  - Talking about methods of attempting suicide, even if not planning
  - Preparation, actually taking actions to prepare for an attempt
  - Having access to lethal means (a way to take their own life), especially firearms

[7] In fact, reliability of clinical assessment for suicidality is a real problem even when direct contact with the patient is available: clinicians are often using some kind of structured interview but also going on instinct, with attendant risks of bias, and most clinicians have not had specialized training for dealing with high risk populations, many of whom are underserved and with special characteristics such as veterans or substance abusers (R. Resnik, 2016).

  - Having enough privacy or isolation to make an attempt

- Context
  - Previous attempts
  - An event or life change that is leading them think about suicide
  - Isolation from friends and family

2.3 Project dataset

Reddit is an anonymous social media site (`http://en.wikipedia.org/wiki/Anonymous_social_media`) in which anonymous users submit posts to areas of interest called 'subreddits'. The labeled dataset we're working with was built using a method inspired by Coppersmith et al. (2014) in their groundbreaking work on quantifying mental health signals in Twitter. The key innovation in that paper was a process for collecting data from people who are likely to have a mental health diagnosis of interest, such as depression:

1. In the first step of their process, a large body of tweets is searched for users who have displayed a clear signal that they *might* have the relevant diagnosis X, namely some regexp-based variant of "I have been diagnosed with X".
2. Second, the resulting "signal" tweet is filtered manually, by deciding whether or not the signal statement was genuine, e.g. filtering out jokes or quotes. ("The Red Sox lost the game. I've been diagnosed with depression!")
3. Users remaining after this filter are considered "positive" instances of the diagnosis, an equal number of "control" users are sampled at random, to be treated as "negative" instances.[8]
4. Having identified a set of positive and negative users, Coppersmith et al. (2014) then collected all the publicly available Tweets for all those users, resulting in a labeled user-level dataset. They then used the resulting data to explore supervised classification of users into positive or negative categories for depression.

In this project, postings to /r/SuicideWatch should be excluded as evidence about a user's suicide risk, since we are focused on screening (identifying if someone should be evaluated as a possible suicide risk) rather than evaluating a level of risk for someone that we already have identified as requiring assessment.

In addition, you should also exclude data from other forums related to mental health, since, even if evidence from those forums proved highly predictive, the content there is specifically generated by people talking about their mental health issues and results would not be generalizable to social media outside of Reddit. The set of other mental health subreddits to exclude includes Anger, BPD, EatingDisorders, MMFB, StopSelfHarm, SuicideWatch, addiction, alcoholism, depression, feelgood, getting_over_it, hardshipmates, mentalhealth, psychoticreddit, ptsd, rapecounseling, schizophrenia, socialanxiety, survivorsofabuse, and traumatoolbox.

Here we talk about the properties of the dataset and how we generate it. The dataset itself contains README files that document the data format.

[8] As a clear limitation of the approach, it must be recognized the self-expression of a diagnosis does not guarantee an actual diagnosis is present. Conversely, a person may have a diagnosis but never say so on Twitter. Nonetheless, this work was a significant step forward in obtaining at least an approximation to ground truth data, given how difficult it is to obtain clinically relevant data for research purposes.

Generating candidates: the full dataset. We've created the dataset inspired by the method of Coppersmith et al. (2014), focusing on suicidality as the mental health condition of interest. We began by identifying users of interest using a July 2015 snapshot of every publicly available Reddit posting as of that date, approximately 1.7B in all.[9] This was done as follows:

1. Rather than looking for a self-report, the "signal" for a potentially positive user is the user making any post to the /r/SuicideWatch subreddit.
2. From the same dataset, we also selected a comparable number of randomly selected control users.
3. We then used the Reddit API to collect all posts by these users from their earliest post up until 2016-07-14. We filtered to include only users who have posted at least 10 posts across all of Reddit.

The resulting dataset contains 11,129 users, with a total of 1,556,194 posts for a total of 81,101,356 words. The candidate, or potentially positive, users are equivalent to the users picked out in the first step of the Coppersmith et al. (2014) process: people who are candidates of positive instances, but who have not yet been reviewed to categorize whether they should truly be considered a positive instance.

Filtering: Crowdsourced judgments. For a subset of 934 randomly selected positive candidates (users) in the full dataset, we crowdsourced an assessment of their degree of risk based on the signal, i.e. their postings to SuicideWatch. In order to facilitate crowdsourced annotation, we divided sequences of more than five SuicideWatch posts for a single user into multiple annotation units containing up to five posts each, yielding a total of 982 annotation units. (For example, a user with 12 posts would yield three annotation units of their first 5 posts, next 5 posts, final 2 posts.) In order to determine user-level risk, we consider a user to have the highest risk associated with any of their annotation units.

We defined a four-way categorization of risk adapting Corbitt-Hall et al. (2016), who provided lay definitions based on risk categories in Joiner Jr et al. (1999):

(a) **No Risk (or "None")**: I don't see evidence that this person is at risk for suicide;

(b) **Low Risk**: There may be some factors here that could suggest risk, but I don't really think this person is at much of a risk of suicide;

(c) **Moderate Risk**: I see indications that there could be a genuine risk of this person making a suicide attempt;

(d) **Severe Risk**: I believe this person is at high risk of attempting suicide in the near future.

These correspond roughly to the *green*, *amber*, *red*, and *crisis* categories defined by Milne et al. in CLPsych ReachOut shared tasks (Milne et al., 2016; Milne, 2017). This process is the equivalent of the second step in the Coppersmith et al. (2014) process, although because of cost we only did it for a subset of 934 candidates rather than all candidates. For purposes of this project, we will consider a user to be a true positive example in this project if their consensus label is c or d.[10]

[9] `https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/`

[10] Optionally, you might want to also experiment with considering only category d as positive instances, since severe risk is often clearest and also highest priority with regard to intervention.

Test set: expert judgments. We selected 245 users at random from the crowdsourced subset to create a set of 250 annotation units that were labeled independently by four volunteer experts in assessment of suicide risk. These included a suicide prevention coordinator for the Veteran's Administration; a co-chair of the National Suicide Prevention Lifelines Standards, Training and Practices Sub-Committee; a doctoral student with expert training in suicide assessment and treatment whose research is focused on suicidality among minority youth; and a clinician in the Department of Emergency Psychiatry at Boston Childrens Hospital. Two of these experts received long detailed instructions laying out risk factors to consider, and the other two were given short instructions telling them to rely upon their experience and training.

Table 1 shows Krippendorff's α — a standard measure of chance-corrected agreement — pairwise among the experts, indicating the set of instructions they used as (S)hort or (L)ong. The average of 0.814 satisfies the conventional reliability cutoff for chance-corrected agreement (> 0.8, Krippendorff (2004)), which is to our knowledge the first result demonstrating inter-rater reliability by clinical experts for suicide risk based on social media postings. Inter-rater reliability for the pair receiving short instructions was substantially lower (0.771), demonstrating the value of our detailed rubric based on explicitly identified risk factors.

We generated consensus user-level labels based on the expert annotations using a generative model for the Dawid-Skene method (Dawid and Skene, 1979; Passonneau and Carpenter, 2014), including consensus for the pairs receiving long instructions (*Long Experts*), short instructions (*Short Experts*), and consensus among all four experts.

Krippendorff α	exp_L1	exp_L2	exp_S1	exp_S2
exp_L1	1	0.838	0.804	0.824
exp_L2	-	1	0.811	0.834
exp_S1	-	-	1	0.771
exp_S2	-	-	-	1

Table 1: Krippendorff's α pairwise among experts

Ultimately this expert-assessed dataset is the true test set we care about: we'd like to be able to identify users in this dataset who are at risk, based on their postings to *non*-mental health forums. As above, we consider a user to be a true positive example in this test set if their consensus label is c or d.[11] *You should avoid looking at the contents of the test set and you should not use it in your exploratory analysis.*

2.4 Additional data

Although it may or may not be useful for this project, we also provide a subset of data from the MyPersonality project,[12] which has collected a very large, anonymized dataset of naturally occurring social data media text data together with personality and in some cases clinical measurements. They did this by creating a Facebook app that allowed people to fill out various kinds of clinical instruments (e.g. questionnaire-based assessments for IQ, Big-5 personality traits such as neuroticism (emotional instability, John and Srivastava (1999)), or depression. People filled out the questionnaires as a fun Facebook app activity (e.g. "how does your assessment of your own personality compare to what your friends say?"), and in the process they would opt in to having their free-text Facebook status updates collected. This produced a collection of datasets involving more than 100,000 people and more than 22 million status updates.[13]

[11] And again, optionally, you might want to also experiment with considering only category d as positive instances, since severe risk is often clearest and also highest priority with regard to intervention.

[12] `mypersonality.org`

[13] If this sounds familiar to you, it might be because it was the work that inspired the approach Cambridge Analytica took to collecting and classifying Facebook user data for purposes of political targeting during the 2016 U.S. presidential election. The people behind Cambridge Analytica took this a step further by acquiring not only the individual's data, but the data for all of their Facebook friends, in violation of Facebook's terms of service. As of April 5, 2018, the current estimate is that data from 87 million people was improperly shared (`https://www.npr.org/sections/thetwo-way/2018/04/04/599542151/`

The subset of MyPersonality data is being made available for this project with permission of the researchers who created the dataset. It includes one subset of data where users filled out a survey used for assessing depression, and another subset where users filled out a survey for a personality inventory involving the traits of openness, conscientiousness, extraversion, agreeableness, and neuroticism.[14] The last of these, neuroticism, is a predictor of depression.[15]

3 The project problem

There are two components to this project: exploratory data analysis using computational linguistics methods and models, and supervised learning to distinguish "positive" instances of suicidality from control users. *Both components of the project are required.*

3.1 Exploratory data analysis

The first goal of this project is to go deeper than you have so far with techniques for exploring differences in language use. Are there detectable differences in the language of the "positive users" (people labeled in categories c or d in the crowdsource-filtered dataset) compared to control users (which come from the unfiltered dataset)?[16]

To tackle this, you should look identify features of language that you think might be worth exploring in social media for identfying positive users, and formulate ideas for how to implement the relevant analysis. Here are just a few ideas, but you should consider these simply as examples and generate ideas of your own also. *You should definitely look at some of the relevant literature that I cited for ideas.*

Operationalizing features that are specifically related to suicide risk. This is probably the most interesting avenue to pursue in terms of doing something new and interesting, although other features below are important to consider also because (a) you'll want to look at the value of interesting features compared to less interesting baselines, and (b) implementing less interesting features that you understand well is a good way of making sure your code is correct. As noted above, Vioulès et al. (2018) describe a number of risk factors and warning signs identified by the American Psychiatric Association and the American Association of Suicidology, and Yates et al. (2017) got excellent results so it's worth looking at their approach to see if it might yield interesting data for exploratory analysis. In addition, our long instructions include a number of factors identified by suicide experts as relevant that one might consider trying to capture through an analysis of posting language and/or user behavior.[17]

Word-based techniques. A baseline approach to any language-based classification task is to look at surface language use, e.g. using simple unigram or n-gram features or association-based methods like the ones we exercised in the homework assignments. It's possible that n-gram language models could potentially pick up differences in use compared to typical language use. Some of the relevant background papers discuss specific sets of words associated with depression or suicidality; it would be interesting to compare what they found with your findings on this dataset.

`facebook-says-cambridge-analytica-data-grab-may-be-much-bigger-than-first-report`). Did I mention that there's an ethical issues section later on in this document that you should read?

[14] `https://en.wikipedia.org/wiki/Big_Five_personality_traits`

[15] https://www.psychologicalscience.org/publications/observer/obsonline/neuroticism-predicts-anxiety-and-depression-disorders.html

[16] A relevant recent paper on this topic is Coppersmith et al. (2016).

[17] We can also give you the full instructions to look at.

VEGETATIVE/ENERGY LEVEL	sleep tired night bed morning class early tomorrow wake late asleep long hours day sleeping nap today fall stay time
SOMATIC	hurts sick eyes hurt cold head tired back nose itches hate stop starting water neck hand stomach feels kind sore
NEGATIVE/TROUBLE COPING	don('t) hate doesn care didn('t) understand anymore feel isn('t) stupid make won('t) wouldn talk scared wanted wrong mad stop shouldn('t)
ANGER/FRUSTRATION	hate damn stupid sucks hell shit crap man ass god don blah thing bad suck doesn fucking fuck freaking real
HOMESICKNESS	home miss friends back school family weekend austin parents college mom lot boyfriend left houston visit weeks wait high homesick
EMOTIONAL STRESS	feel feeling thinking makes make felt feels things nervous scared lonely feelings afraid moment happy worry comfortable stress excited guilty
ANXIETY	feel happy things lot sad good makes bad make hard mind happen crazy cry day worry times talk great wanted

Table 2: LDA-induced themes related to depression.

Word classes. Lexical techniques can be extended to consider word *categories*, rather than just words — for example, Pennebaker's Linguistic Inquiry and Word Count dictionary (LIWC, Pennebaker and King (1999)) makes it possible to look at word categories like NEGEMO (negative emotion words) or INSIGHT (including words like *accept*, *admit*, *believe*, *conclusion*, *explanation*).[18] It's also possible that WordNet categories could be useful here.[19] See also Empath, another approach to lexical categories; some of their categories (e.g. love, sympathy, irritability, nervousness, etc.) may be particularly relevant.[20]

Topic models. Topic models provide fairly general way of capturing thematic content or trends, which can also be relevant in many language classification tasks. As an example in mental health, Resnik et al. (2013) looked at emotional instability and depression using topic models in a corpus of writing by college students (Pennebaker and King, 1999). Table 2 shows seven topics identified by a clinician as particularly indicative of potential depression and individuals meriting further evaluation. These induced topics capture problem-specific and even population-specific properties in ways that *a priori* lexical resources cannot — for example, although the widely used Linguistic Inquiry and Word Count lexicon has a *body* category, it does not have a category that corresponds to somatic complaints, which often co-occur with depression. Similarly, some words related to energy level, e.g. *tired*, would be captured in LIWC's *body*, *bio*, and/or *health* category, but the LDA theme corresponding to low energy or lack of sleep, another potential depression cue, contains words that make sense there only in context (e.g. *tomorrow*, *late*). Other themes, such as the one labeled HOMESICKNESS, are clearly relevant for depression (potentially indicative of an adjustment disorder), but even more specific to the student population and context. It's not clear that topical or thematic distinctions like these are relevant for the present task, but it is worth considering.

Table 3 illustrates topics obtained by running a 50-topic *supervised* topic model (sLDA) on the Pennebaker stream-of-consciousness dataset (Resnik et al., 2015). This analysis used, as each essay's regression variable, the student's degree of neuroticism — a personality trait that can be a risk factor for internalizing disorders such as depression and anxiety — as assessed using the Big-5 personality inventory (John and Srivastava, 1999). The neuroticism scores are Z-score normalized, so the more positive (negative) a topic's regression value, the more (less) the supervised model associates the topic with neuroticism. A clinician identified the most relevant topics; these were presented in random order without the neuroticism regression values in order to avoid biasing the judgments. The sLDA neuroticism values for topics in Table 3 pattern nicely with the clinician judgments: negative neuroticism scores are associated with clinician-judged positive valence topics, and positive neuroticism scores with negative valence. Scores for the p and n valence items differ significantly according to a Mann-Whitney U test ($p < .005$).

[18] As a pointer to work on a different category of mental disorder, Fineberg et al. (2015) use LIWC to explore differences in word class use associated with schizophrenia; see also Hong et al. (2012).

[19] `https://wordnet.princeton.edu/`

[20] `http://empath.stanford.edu/`

Notes	Valence	Regression value	Top 20 words
social engagement	p	-1.593	game play football team watch win sport ticket texas season practice run basketball lose soccer player beat start tennis ball
social engagement	p	-1.122	music song listen play band sing hear sound guitar change remind cool rock concert voice radio favorite awesome lyric ipod
social engagement	p	-0.89	party night girl time fun sorority meet school house tonight lot rush drink excite fraternity pledge class frat hard decide
social engagement	p	-0.694	god die church happen day death lose doe bring care pray live plan close christian control free hold lord amaze
high emotional valence	e	-0.507	hope doe time bad wait glad nice happy worry guess lot fun forget bet easy finally suck fine cat busy
somatic complaints	n	-0.205	cold hot hair itch air light foot nose walk sit hear eye rain nice sound smell freeze weather sore leg
poor ego control; immature	n	0.177	yeah wow minute haha type funny suck hmm guess blah bore gosh ugh stupid bad lol hey stop hmmm stuff
relationship issues	n	0.234	call talk miss phone hope mom mad love stop tonight glad dad weird stupid matt email anymore bad john hate
homesick; emotional distress	n	0.34	home miss friend school family leave weekend mom college feel parent austin stay visit lot close hard boyfriend homesick excite
social engagement	p	0.51	friend people meet lot hang roommate join college nice fun club organization stay social totally enjoy fit dorm conversation time
negative affect*	n	0.663	suck damn stupid hate hell drink shit fuck doe crap smoke piss bad kid drug freak screw crazy break bitch
high emotional valence	e	0.683	life change live person future dream realize mind situation learn goal grow time past enjoy happen control chance decision fear
sleep disturbance*	n	0.719	sleep night tire wake morning bed day hour late class asleep fall stay nap tomorrow leave mate study sleepy awake
high emotional valence	e	0.726	love life happy person heart cry sad day feel world hard scar perfect feeling smile care strong wonderful beautiful true
memories	n	0.782	weird talk doe dog crazy time sad stuff funny haven happen bad remember day hate lot scar guess mad night
somatic complaints*	n	0.805	hurt type head stop eye hand start tire feel time finger arm neck move chair stomach bother run shoulder pain
anxiety*	n	1.111	feel worry stress study time hard lot relax nervous test focus school anxious concentrate pressure harder extremely constantly difficult overwhelm
emotional discomfort	n	1.591	feel time reason depress moment bad change comfortable wrong lonely feeling idea lose guilty emotion confuse realize top comfort happen
homesick; emotional distress*	n	2.307	hate doe sick feel bad hurt wrong care happen mess horrible stupid mad leave worse anymore hard deal cry suppose

Table 3: sLDA topics from Pennebaker stream-of-consciousness essays identified by a clinician as most relevant for assessing depression. Supervision (regression) is based on Z-scored Big-5 neuroticism scores.

Readability measures. De Choudury et al. discuss readability as a potential source of signal for whether a user on a Reddit mental health forum is likely to later post on SuicideWatch. Various standard measures of "readability" capture lexical and/or syntactic factors. See, e.g., `https://pypi.python.org/pypi/readability/0.1`.

Syntactic variation. Another intriguing possibility to consider is that variation in *syntactic* choices might be related to underlying mental health status. It is well known from the lexical semantics literature that grammatical constructions are linked to underlying semantic properties such as causation (was an event caused or did it just happen?), volition (did the agent of the event intend to make it happen?), telicity (did the event have a defined endpoint?), and affectedness (was the object of an event affected by it?). Greene and Resnik 2009 showed that these semantic properties can mediate the relationship between what people hear and their judgments based on what they hear — for example, given a story about an event where somebody kills someone else by drowning them, a headline like *Victim drowns* is perceived as more sympathetic to the perpetrator than *Perpetrator Drowns Victim*, because, in contrast to a subject-verb-object transitive structure, an inchoative construction like *Victim drowns* de-emphasizes the causal and volitional role of the perpetrator and the affectedness of the victim. As a real-world example, when the chairman of British Petroleum testified in front of the U.S. Congress about the Deepwater Horizon oil rig disaster, he referred to "an explosition in which eleven workers were lost", not an explosion that killed eleven workers.[21]

How might syntactic variation be related to depression or suicidality? One could use computational linguistics methods to explore a number of hypotheses related to the concept of negative attentional bias, that is, the finding that people suffering from depression tend to focus more on negative information (Feng et al., 2015).

[21] Verbs involving killing are linguistically well suited for these discussions because *kill* and similar verbs are canonically Transitive semantically, i.e. a killing event canonically involves causation, volition, affectedness of the object, a defined endpoint, etc.; see Greene and Resnik for discussion. As a less grisly example, my favorite manifestation of this kind of "syntactic framing" is when my 6-year-old says "Daddy, my toy broke" (inchoative) instead of "Daddy, I broke my toy" (transitive).

For example, one hypothesis might be that, beyond simply using more negative words (which is already well established), someone who is depressed might be more likely to put themselves as the *object* of a negative verb, consistent with the perception of being affected by negative states or events. Conversely, one might hypothesize that a depressed person might be *less* likely to view themselves as capable of causally affecting things around them in a positive way, and therefore less likely to use language where they are the agent of a positive, causal event. Pennebaker has found predictive differences in pronoun use: depressed people use the word I much more often than emotionally stable people, likely reflecting an inward-facing perspective; but of course that pronoun in English only appears in subject position, so could there be something deeper going on that involves not only the subject but also the syntactic constructions and/or the positive-or-negative valence of the verb? Taking this a step further, perhaps similar distinctions in viewpoint might exist more generally whether or not the person himself or herself is involved in the event, e.g. a greater use of detransitivizing constructions (inchoative, passive) might be connected with a general view of the world as involving things that "just happen" as opposed to being caused with a purpose.

Other forms of dimensionality reduction. Bedi et al. (2015) use latent semantic analysis (LSA) as a way to capture semantic content, as a way of operationalizing the idea that people suffering schizophrenia often manifest greater discontinuity of thought, e.g. "derailment", where someone's language includes sequences of unrelated or only remotely related ideas. Along with LDA, LSA or deep learning techniques could be used to explore lower-dimensional lexical, sentence, or document representations, and/or semantic trajectories or consistency of content within or across posts. The latter point also raises the possibility that other sequential characteristics of the language might be relevant.

Non-language measures. The main focus for this project, obviously, is natural language processing. But it would be perfectly reasonable to also explore some non-language characteristics such as average volume of postings, lengths of postings, or temporal patterns in postings such as whether people are more likely to be posting late at night (e.g. bucketing timestamps into 3- or 4-hour windows). One could also combine that with language characteristics, too, e.g. perhaps symptom domains such as agitation can be detected from language but are more relevant when they're seen very late at night.

These are just a few ideas — you should look at relevant papers and it's likely you'll also come up with others!

Once you've got a set of features that you hypothesize might be useful, there are a number of ways you might consider exploring them in the data. Statistical hypothesis testing is certainly one: for any given feature, you could evaluate the hypothesis that it appears among positive users more often than among control users. This is also a way of doing feature selection for supervised learning (see e.g. `http://scikit-learn.org/stable/modules/feature_selection.html`). Another possibility would be using principal components analysis (PCA) to take a larger set of features and reduce it to (hopefully) interpretable subsets. Still another would be to take a representation learning approach to see whether a network could learn higher-level abstract features that capture information relevant to the task. And yet another after that might be to use attention in a neural network to highight parts of postings that are particularly relevant; in this domain, a nice example is the explanation generation approach in Kshirsagar et al. (2017).

Your assignments have included examples of potential outcomes of exploratory data analysis, e.g. hypothesis tests, top-N features that distinguish among the groups of interest, or heat maps or other visualizations that might help bring interesting patterns to the surface.

Note again that exploratory data analysis must exclude test data, to avoid coming up with ideas for features that are overfitted to the test set.

3.2 Supervised classification

The second goal of this project is to explore a supervised learning approach to distinguishing "positive" users from controls, using linguistic and possibly other features. Classification should be evaluated using precision, recall, and F1-measure on the held-out test set (245 positives and 245 controls).

Note that precision (and therefore also F-measure) is sensitive to the distribution of data in the test set, which means that performance on a balanced test set doesn't really give you a good indication of how your technique would perform in the real world. (Positives are hugely over-represented in the test set compared to the real world – thankfully! – which means that if your classifier chooses positive when it shouldn't, it has a better chance of being right in this test set. Since false positives aren't being penalized as much as they should be, precision is overestimated.) A good solution to this problem is to also generate receiver-operating characteristic (ROC) curves, and to use ROC AUC (area under the curve) as a summary (scalar) evaluation measure rather than F1 `http://en.wikipedia.org/wiki/Receiver_operating_characteristic`. I recommend doing this also, although it's optional.

All of the possibilities in Section 3.1 are certainly fair game in terms of features, and you can propose more if you like. You should follow the recommendations of Resnik and Lin (2010) when it comes to supervised learning methodology, including, for example, evaluating improvements against lower-bound baseline (such as unigram features), keeping training and test data separate, etc.

One key decision to make is whether to use all of the data from a user as a single training instance ("document"), or whether to do something more sophisticated in terms of modeling. Here are two things to consider:

- You have postings from users that have been made over a period of time, and suicidality is a property of a person at a particular point in time — as opposed, for example, to personality traits, which tend to remain relatively stable over time. Grouping together all of a person's postings may be too coarse grained.
- The fact that a user is a positive instance certainly does not mean that *everything* they post reflects their suicidality.

There are a variety of ways one might address these issues, at different levels of ambition/complexity. For example, if the user's first SuicideWatch posting was at time t, one might consider restricting relevant evidence to posts they made within some periods prior to t, e.g. trying a day, two days, a week, a month (addressing the first problem), but still assuming all posts are equally relevant (not addressing the second problem). More ambitiously, one might consider whether suicidality should be treated as a latent variable — e.g. in a generative model, the generative story might include a Bernoulli choice as to whether or not a latent positive suicidal state will or will not be reflected in a posting, and then the content of the posting could be generated from different distributions depending on the outcome of that coin flip. (This should sound somewhat similar to the naïve Bayes model from the Gibbs Sampling tutorial.) Even more ambitiously, you might consider sequence modeling for those latent states, neural sequence modeling, an attention model, reinforcement learning... There are lots and lots of possibilities. What's most important is being thoughtful about your choices, demonstrating your understanding/proficiency with material we've covered in the course, and producing a strong, convincing report of what you did.

4 What you need to do

Project plan. This project is deliberately underspecified: the first part of your job, after figuring out who you'll work with, is to scope out a project that will be feasible within the necessary time frame.

Creating the project plan is your homework assignment for this week. The biggest risk here is carving off a project that is too ambitious to be done with the time you have, so it's *very* important to run the project plan by me, for me to do some reality checking. You should give me a proposed project plan that describes what your group plans to try, in enough detail that I can provide you feedback and guidance, and particularly so that I can steer you away from approaches that are likely to get you bogged down. *In doing this, please organize your propsal according to the outline of the project writeup (below).*

To create your project plan, I strongly recommend that you look at relevant papers to (a) identify relevant properties of language that you're going to explore, and (b) sketch out how you plan to operationalize those properties algorithmically. As part of this process, I recommend taking a look at the dataset, in this process. (To be pure, don't look at test data.)

You don't need absolutely every detail worked out; just give me enough so I can see clearly what you're aiming for. A couple of pages is probably plenty. And if the writeup structure isn't a good match for what you're going to do, that's ok, too, it's fine to diverge from it. What's important is that you're getting an early start on looking at data and thinking through the issues and your plan, and give me a chance to comment. Make sure to leave room for unanticipated problems — messy data that could need to be cleaned up, etc. As I emphasize further below, this is not a textbook exercise; you're playing with real-world stuff, and real-world problems are unpredictable.

You are more than welcome to use/adapt off-the-shelf code rather than implementing things yourselves. In fact, *this is strongly encouraged*. I want you to spend your time *exercising* what you've learned, not creating your own implementations of SVM classification, LDA, syntactic parsing, etc. I can provide some pointers to things, and you can also use the class discussion board to talk about code, implementations, etc. *No group will be penalized for intellectual generosity in sharing what they learn with other groups!* Please just make sure to acknowledge others in your writeup if they were helpful to your group, saying explicitly what they did that helped.

Project deliverables. Here's what we expect to be turned in by your group on the due date.

1. A tarball/zipfile of your source code, *including enough information for someone to run it without having to ask you questions*. This should include a README that walks the reader through what to type on the command line(s) to use your code. You needn't include software components that someone can download, (e.g. WEKA), but please include all necessary information, e.g. your README might say something like *Change the value of the variable weka-dir in file run-training to the directory in which you installed WEKA*. (Good practice, although not required, would be to actually provide an executable build script that downloads the relevant stuff, e.g. commands like *pip install spacy* (python) or *cpanm Module::Name* (perl).)

   If you prefer to provide a link to a code repository instead that's fine, as long as the README.md fits the above requirements. Please just remember that the datasets themselves or other resources (e.g. the LIWC lexicon) absolutely *cannot* be put anywhere publicly accessible.

2. A writeup with at least the elements below. Please stick with this main structure, though you can add sections if you need to. If for some reason you feel you need to depart significantly from this, please contact me in advance to discuss.

   Your group's grade will be based primarily on the writeup. I can't stress this enough. The writeup is important, because it's the main thing I'm looking at, so make sure you produce something clean and well written and budget in a lot of time so you can do it well. As an important note, a really common problem I've seen is groups breaking the writeup into sections, giving each person a section, and then simply throwing them together at the end. This results in really uneven quality and writeups that are often quite hard to read. There's a difference between a group effort and a union (or concatenation!)

of individual efforts. Aim for the former, not the latter. **If you have to scale back how ambitious your project is in order to ensure a good writeup, that's the right choice to make.**

Take a look at recent papers I've cited on computational linguistics for mental health (e.g. Yates et al. (2017)) as examples of well written papers in this subject area.

If you would like early feedback on a draft writeup, you're welcome to send one to me. Please just leave lots of time for me to read it and for you to revise things based on my comments.

Here is the basic structure for the writeup:

(a) Introduction

i. Who is in this group. Also provide a rough breakdown of people's roles in the project. If roles weren't broken out cleanly, that's ok, identify people's roles as clearly as possible. I'm mainly just interested in how you organized things and in seeing that everyone contributed.

ii. High level description of what you decided to do and why, and what you expected (or at least hoped) to get out of it. Although in principle it would be good for you to get practice at providing a motivating introduction like I gave in Section 1, the way that people do in a conference or journal paper, you do *not* need to do so, unless you've got something new to say that hasn't been said above or in previous literature. I already know what I said and I've already read a lot of the previous literature, and I would much rather you spend your time on the sections of the writeup that really matter. Definitely do not just spit back material from this document.

(b) Data and Methods

i. Any data and resources you used other than what I've already given you. (For that you can just say you used the data and resources that you were given for the project.) For any other data or resources: how you got it, basic properties (size, etc.),. If applicable, include anything you needed to do with data or resources (including what I gave you) in order to work with it.

ii. Basic information about preprocessing, e.g. how you did tokenization, removal of stopwords (if you did that), etc.

iii. Relevant descriptions of which language (and metadata, if applicable) characteristics you looked at and why. Make sure to cite relevant source as appropriate. (Please use a citation style that includes the authors *inline*, e.g. "a fantastic paper (Resnik, 1999)", not "a fantastic paper [13]".).
Include a description of what you did to operationalize or implement the text analysis to capture those characteristics.You do *not* need to regurgitate textbook- or article-style descriptions of existing algorithms, just point to the source (bibliographic reference and, if relevant, where you got code), if you are using something that exists rather than designing something new. Again, note that you are *not* required to invent new things or implement from scratch for this project; applying what you've learned to this new problem space is fine. However, you *do* need to provide relevant details. As a good example, consider something like this excerpt (made up for purposes of illustration):

> "To obtain word classes based on topic modeling, we trained Chang's implementation of sLDA (Blei et al. 2008, `http://cran.r-project.org/web/packages/lda`)) with 40 topics, using each author's combined set of posts as the document, and that author's group (+1 for positive, -1 for control) as the response variable. We chose $k = 40$ as the number of topics by trying values between 20 and 50 to see which worked best on held-out (dev) data. See Table 3 for the 40 topics, and see Appendix A for excerpts from of several documents along with the posterior distribution of topics for each example document."

(c) Relevant information about any other algorithms and models, e.g. PCA, supervised classification, etc. Identify what approach you took *and why*, which software you used and its relevant parameters, etc.

(d) Evaluation

i. For exploratory data analysis, present a well structured, informative discussion of what you found (or didn't find), including examples, figures, tables, etc. as appropriate.

ii. For classification, describe how you evaluated what you did in development and final testing, e.g. including details like cross-validation if you used it, evaluation metrics, etc. For discussion of evaluation metrics and presentation of results, see Lin and Resnik *Evaluation in NLP* and good prior papers. As an example of describing development, you might find yourself including a statement like the following:

> "We tuned the α and β parameters using a grid search with values of 0.01, 0.05, and 0.1 for each parameter. For testing, we then used the combination of α and β that performed best in the grid search as evaluated using 5-fold cross validation on the training data."

iii. Analysis and discussion. This should include not just a summary of the quantitative results, but a qualitative look at how things worked. For example:

> "We looked at the features receiving the strongest weights in our trained regression model. The most influential features included LIWC's *negemo* (negative emotion words) normalized frequency; the binary word-is-present features for individual negative emotion words like *bummed* and *sad*; the topic feature (posterior topic weight) for topics 12 in Table 3, which seems to capture language pertaining to anger/frustration; and the 'topical coherence' feature we implemented based on Bedi et al. (2015) (discussed above in Section 2.2)."

iv. Ethical issues. Read Benton et al., "Ethical Research Protocols for Social Media Health Research", `http://www.ethicsinnlp.org/workshop/EthNLP-2017.pdf#page=106` and discuss each of the issues in Sections 3.1 through 3.8 in terms of what you did or did not do in this project (or, if it's not relevant, explain why). A sentence or two is fine for these — this doesn't need to be a focus of the writeup or take a lot of time, but I do want each group to have read and discussed this.

(e) Discussion and future work

i. Qualitative discussion and conclusions. In what ways did you succeed, and in which ways didn't you? Are there any surprises in the data, or interesting things to highlight — more generally, what did you learn? What directions seem most promising for future work?

ii. Optionally, any particular difficulties or hurdles you encountered. Please feel free to include ways in which final projects like this could be made better.

(f) Separately, by e-mail, each person should send to the TA e-mail address three ratings for their team members as described under "Grading", below. Please put "Compling2 ratings - YOUR NAME" in the subject line so these messages are easy to spot. Attach a file named `YOUR_NAME.ratings.tsv` containing the following (tab-separated) *in exactly this format*:

```
your_name    teammate1_name    rating    Justification
your_name    teammate2_name    rating    Justification
```

Yes, we want your name repeated, identically, in the first column. Please agree within your team on a consistent way to write everyone's name. This way we can concatenate all the files we receive and sort by the 2nd column to easily aggregate the ratings for each person.

5 Grading

The group will receive a grade-in-common out of 90 points. By now I think you have a decent sense of my criteria, and I've been pretty explicit above. Thoughtfulness, effort, looking at data and output/results to achieve insight, and exercising things you've learned in class this semester (not just things you already knew) —these are the things I care most about, and the only way you can communicate them effectively is through a carefully thought-through, well written writeup.

> *Note that late projects will not be accepted. If a project is turned in late, the group project score will be zero. Budget your time accordingly and do not leave anything to the last minute.* If you have an emergency of some kind, let me know as soon as it becomes an issue, not afterward. To repeat what I say on the course homepage, emergencies include urgent medical issues, family emergencies, or other valid reasons we can discuss if necessary. What's crucial is that if you do have a problem or issue, you talk to me about it as soon as possible. I can tell you in advance that there are several common problems I absolutely will not consider as valid reasons for failing to get work in on time. These include (a) failure to manage your time properly, including non-emergency travel, being busy with another course, a piece of research, a conference presentation, a paper submission deadline, etc.; (b) discovering an assignment is harder or takes longer than you expected it to (see item a); and (c) losing code or data that should have been backed up, unless it's someone else's fault (e.g. you backed things up on a department server and it failed).

For the remaining 10 points for each individual, each team member will anonymously rate each other team member on a scale from 1 to 10, taking into account three criteria: *collaboration* (would you like to work with this person again, or were they hard to work with?), *contribution* (did this person contribute their expected or fair share to the project, where "fair" includes *adjusting* for the fact that different people have different abilities?), and *effort* (did this person try really hard, whether or not they wound up succeeding in contributing well)? Your credit for these 10 points will be defined by the average of the ratings you receive, modulo my discretion after looking at the justifications. I expect that in most groups, everyone will probably get 10 out of 10, and that's fine. The point of these ratings is to encourage people to work well together — if you don't do right by your teammates, it should show up in your individual grade. Note that for any rating less than 10, the *justification* column should contain a statement justifying the score in detail, e.g.: *I gave Philip a 5 because he ended up contributing a lot less than we all had agreed on. Specifically, his main expected piece of the project was to explore RNNs and autoencoders for generating vector representations. Although he did an ok job on the RNNs, he basically disappeared after the first two or three weeks, not coming through on the autoencoder piece and, more important, barely contributing to the writeup.*

6 Ethical use of data

6.1 General notes

As noted above, you need to read Benton et al. (2017). Most of what I'm saying here is covered there, and more.

Human subjects research, which is overseen by the university's Institutional Review Board, is defined as (a) a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge, that involves (b) a living individual about whom a research investigator obtains data through intervention or interaction with the individual, or individually identifiable information.

Student class assignments like this one are not generally considered human subjects research, for two reasons. First, they are not research, because they are intended to help train students or give them experience with research methods, as opposed to collecting information systematically with the intent to develop or contribute to generalizable knowledge. (The intent matters: I hope you'll learn enough from this assignment to be able to *do* good research, possibly even to follow up this class assignment with a real research project — see below — but the work you're doing for this project during this semester is not intended to produce publications.) In addition, assignments like this one don't involve "human subjects", technically speaking: we are working with publicly available social media behavior, which involves neither intervention, nor interaction with individuals, nor individually identifiable information, particularly since Reddit is an anonymous social media site.

That said, any project involving social media needs to be handled with great sensitivity, particularly when touchy issues like mental health are involved. It is important, therefore, that you not disseminate or share the data we are working with, and it would also be completely inappropriate to use Web searches to look for further information from or about a user in these datasets, even for benign purposes. Following Benton et al. (2017), rather than quoting any individual postings in your writeup, you should carefully paraphrase, so that someone else doing a search would be less likely to find the posting.

If you have any questions about appropriate use of data, please let me know.

I would add that in any study involving naturally occurring, real-world data, it is possible that you will come across material that you might consider inappropriate, obscene, or upsetting — no greater or less than what you would usually encounter in daily life. If this is a concern for you, also please let me know. Although the data you're getting for the project includes SuicideWatch postings, and therefore I felt it important to include the big warning on the first page, one of the reasons I chose screening (based on people's everyday postings) rather than risk assessment (based on their SuicideWatch postings) as the problem was to minimize exposure to those posts.

If you are interested in further discussion about the ethics of research on social media, here are a few more sources.

- Mikal et al., "Ethical issues in using Twitter for population-level depression monitoring: a qualitative study", `https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-016-0105-5`
- Solberg (2012)
- Michelle Meyer's blog post, "How an IRB Could Have Legitimately Approved the Facebook Experimentand Why that May Be a Good Thing" (http://www.thefacultylounge.org/2014/06/how-an-irb-could-have-legitimately-approved-the-facebook-experimentand-why-that- may-be-a-good-thing.html)

6.2 Use of the dataset

The primary dataset for this educational project has been collected from an online social media source. The following specifies conditions for proper use of the dataset.

1. Privacy of the users and their data is critical. Absolutely no attempt can be made to de-anonymize or interact in any way with users.
2. This project is being done solely for educational purposes, and your results cannot be used directly in research papers. If you get promising results and would like to develop the ideas into a research paper for publication, or to use what you've done further for another class, please talk with me about obtaining suitable Institutional Review Board approvals.

3. You may not use these data for any purpose other than this specific class project. You should not show or share this data with anyone outside class, and you may not do any research or development on this dataset outside the scope of the class project. If there are things you're interested in doing with this dataset outside the scope of the class project, please talk with me.

4. Once you have completed the project, you are expected to delete any copy of the dataset you have made, including any derived files (e.g. tokenized versions of the documents). It is ok to keep the results of feature extraction as long as the original text cannot be reconstructed from that data.

5. You *should not* cut/paste any text content from this dataset into your proposal, your paper, onto the class discussion board, into e-mail, etc. If you want to identify a specific posting, e.g. in discussion on the class discussion board, use the ID from the dataset. If you want to give examples, please create a paraphrase instead of the original text. For example, if a posting said *What's this world come to? http://t.co/ XxI4QnMew* you could change it to *I wonder what this world has come to? http://t.co/ YYY.* (Or just make up a post that demonstrates whatever it is you want to describe.)

In both your project plan and your final project writeup, please include the following statement: *We have read and understood the conditions on proper use of the project dataset.*

Unless you speak with me in advance about keeping the data for collaborative research, in your final project writeup please also include the following statement: *We have deleted all our copies of the project dataset.*

If you have any questions or concerns, of course please speak with me.

7 A Final Note

This project is ambitious. *Really* ambitious. It attempts to give you an experience doing something real, not just a textbook exercise. It's an extension of things we've done before, but it's also the first time we're trying this specific project formulation. That means that there might be unanticipated problems, situations where people do not receive inputs they need to get their part done, intra-team politics, interpersonal issues, and who knows what else — just like in the real world. It also means that what you do for learning purposes here could wind up sparking some new and interesting ideas for a real research project to follow, which is pretty cool.

Unlike the real world, which is not very forgiving, this is a controlled setting that involves the guidance of an instructor, who *can* be very forgiving. Remember that the activity is, first and foremost, a collaborative learning activity, with the emphasis on *learning.* If there are problems or issues of any kind, let me know as soon as possible, and I will help to get them worked out. Also feel free to use the class discussion forum: to emphasize again, the presence of multiple teams does *not* mean that you are competing with each other. Contributing to the class as a whole earns extra points, in my book.

And remember to have fun!

References

APA. 2013. The critical need for psychologists in rural america. Http://www.apa.org/about/gr/education/rural-need.aspx, Downloaded September 16, 2013.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders, 5th edition*. American Psychiatric Association.

Philip J Batterham, Maria Ftanou, Jane Pirkis, Jacqueline L Brewer, Andrew J Mackinnon, Annette Beautrais, A Kate Fairweather-Schmidt, and Helen Christensen. 2015. A systematic review and evaluation of measures for suicidal ideation and behaviors in population-based research. *Psychological assessment* 27(2):501.

Gillinder Bedi, Facundo Carrillo, Guillermo A Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B Mota, Sidarta Ribeiro, Daniel C Javitt, Mauro Copelli, and Cheryl M Corcoran. 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia* 1.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. pages 94–102.

Bureau of Health Workforce. 2017. Designated health professional shortage areas: Statistics, first quarter of fiscal year 2018, designated hpsa quarterly summary. Health Resources and Services Administration (HRSA) U.S. Department of Health & Human Services, `https://ersrs.hrsa.gov/ReportServer?/HGDW_Reports/BCD_HPSA/BCD_HPSA_SCR50_Qtr_Smry&rs:Format=PDF`.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering* 23(5):649–685.

Mike Conway and Daniel OConnor. 2016. Social media, big data, and mental health: current advances and ethical implications. *Current opinion in psychology* 9:77–82.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. pages 51–60.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015a. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. North American Chapter of the Association for Computational Linguistics, Denver, Colorado, USA.

Glen Coppersmith, Ryan Leary, Eric Whyne, and Tony Wood. 2015b. Quantifying suicidal ideation via language usage on social media. In *Joint Statistics Meetings Proceedings, Statistical Computing Section, JSM*.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*. pages 106–117.

Darcy J Corbitt-Hall, Jami M Gauthier, Margaret T Davis, and Tracy K Witte. 2016. College students' responses to suicidal content on social networking sites: an examination using a simulated Facebook newsfeed. *Suicide and life-threatening behavior* 46(5):609–624.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics* pages 20–28.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 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*. ACM, pages 2098–2110.

Zhengzhi Feng, Xiaoxia Wang, Keyu Liu, Xiao Liu, Lifei Wang, Xiao Chen, and Qin Dai. 2015. The neural mechanism of negative cognitive bias in major depressiontheoretical and empirical issues. `https://www.intechopen.com/books/major-depressive-disorder-cognitive-and-neurobiological-mechanisms/the-neural-mechanism-of-negative-cognitive-bias-in-major-depression-theoretical-and-empirical-issues`.

SK Fineberg, S Deutsch-Link, M Ichinose, T McGuinness, AJ Bessette, CK Chung, and PR Corlett. 2015. Word use in first-person accounts of schizophrenia. *The British Journal of Psychiatry* 206(1):32–38.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*. Association for Computational Linguistics, pages 503–511.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18:43–49.

Kai Hong, Christian G Kohler, Mary E March, Amber A Parker, and Ani Nenkova. 2012. Lexical differences in autobiographical narratives from schizophrenic patients and healthy controls. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pages 37–47.

Oliver P John and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research* 2:102–138.

Thomas E Joiner Jr, Rheeda L Walker, Jeremy W Pettit, Marisol Perez, and Kelly C Cukrowicz. 2005. Evidence-based assessment of depression in adults. *Psychological Assessment* 17(3):267.

Thomas E Joiner Jr, Rheeda L Walker, M David Rudd, and David A Jobes. 1999. Scientizing and routinizing the assessment of suicidality in outpatient practice. *Professional psychology: Research and practice* 30(5):447.

Klaus Krippendorff. 2004. Reliability in content analysis. *Human communication research* 30(3):411–433.

Rohan Kshirsagar, Robert Morris, and Samuel Bowman. 2017. Detecting and explaining crisis. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, pages 66–73. http://aclweb.org/anthology/W17-3108.

David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics, San Diego, CA, USA, pages 118–127. http://www.aclweb.org/anthology/W16-0312.

D.N. Milne. 2017. Triaging content in online peer-support: an overview of the 2017 CLPsych shared task. Available online at `http://clpsych.org/shared-task-2017`.

Danielle Mowery, Hilary Smith, Tyler Cheney, Greg Stoddard, Glen Coppersmith, Craig Bryan, and Mike Conway. 2017. Understanding depressive symptoms and psychosocial stressors on twitter: A corpus-based study. *Journal of Medical Internet Research* 19(2).

Kevin A Padrez, Lyle Ungar, Hansen Andrew Schwartz, Robert J Smith, Shawndra Hill, Tadas Antanavicius, Dana M Brown, Patrick Crutchley, David A Asch, and Raina M Merchant. 2015. Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department. *BMJ Qual Saf* pages bmjqs–2015.

Rebecca J Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics* 2:311–326.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology* 77(6):1296.

Dina Popovic, Eduard Vieta, Jean-Michel Azorin, Jules Angst, Charles L Bowden, Sergey Mosolov, Allan H Young, and Giulio Perugi. 2015. Suicide attempts in major depressive episode: evidence from the bridge-ii-mix study. *Bipolar disorders* 17(7):795–803.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond LDA: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*.

Philip Resnik, Andy Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. Poster session.

Philip Resnik and Jimmy Lin. 2010. Evaluation of nlp systems. *The handbook of computational linguistics and natural language processing* 57:271.

Philip Resnik, Rebecca Resnik, and Margaret Mitchell, editors. 2014. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Baltimore, Maryland, USA. http://www.aclweb.org/anthology/W/W14/W14-32.

Rebecca Resnik. 2016. Psychological assessment: The not good enough state of the art. Presentation, Veterans Affairs Suicide Prevention Innovations Conference (VASPI).

Kathleen Sibelius. 2013. Increasing access to mental health services. Http://www.whitehouse.gov/blog/2013/04/10/increasing-access-mental-health-services.

Lauren B. Solberg. 2012. Regulating human subjects research in the information age: Data mining on social networking sites. *Northern Kentucky Law Review* 39(2). Http://ssrn.com/abstract=2157302.

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

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 2968–2978. https://www.aclweb.org/anthology/D17-1322.