Computational Linguistics I Final Project Ideas
Idea A: Exploratory Data Analysis for Linguistic Signal for Suicidality in Social Media
Idea B: Predictive Modeling Using 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 an ongoing research 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 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 is available. 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 2, below, on the ethics of working with social media data.

1 Introduction

Let's start with some numbers.

- Combining direct and indirect costs, the global cost of mental health conditions for 2010 was estimated at $2.5 trillion dollars.[1]
- Looking at the economic toll of non-communicable disease, if you look at the projected cost of mental illness worldwide taking indirect costs into account (e.g. not just medical care costs, but also things like lost income), the cost outstrips cardiovascular diseases, and it's more than the costs of diabetes, cancer, and chronic respiratory diseases *combined*.[2]
- In the U.S. more than 115 million people live in federally designated Mental Health Care Professional Health Professional Shortage Areas — that is, they live in places where it's hard to get mental health treatment, even if they realize they need help in the first place, which often they don't.[3]
- Suicide is the third leading cause of death among youths and young adults aged 10 to 24 years, and second for ages 15-19.[4]
- Based on a comprehensive meta-analysis of 365 studies (3,428 total risk factor effect sizes), Franklin et al. (2017) concluded that predictive ability for suicidal thoughts and behavior has not improved across *50 years* of research.

I could go on, but it seems clear that 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;

[1] The Global Economic Burden of Non-Communicable Diseases. World Economic Forum and Harvard School of Public Health, September 2011

[2] `http://www3.weforum.org/docs/WEF_Harvard_HE_GlobalEconomicBurdenNonCommunicableDiseases_2011.pdf`

[3] `https://www.kff.org/other/state-indicator/mental-health-care-health-professional-shortage-areas-hpsas`

[4] Rebecca Volker, Youth Suicide Linked with Economic Downturns, news@JAMA, August 14, 2014.

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

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.[6]

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 (e.g. 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. 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 project focuses on risk assessment 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.

Under normal circumstances this project would have two parts: exploratory data analysis, to dig into the dataset, and then using what you've discovered predictive modeling. Because the COVID-19 crisis is creating significant new challenges for people, however, this year the project will involve doing *either* exploratory data analysis *or* predictive modeling (with error analysis then providing the opportunity to look more closely at the data).

2 Ethical use of data

2.1 General notes

Whenever you're working with data that originates with human beings, it's important to spend some time thinking about appropriate uses of the data both in terms of official rules, and in terms of broader ethical considerations whether or not those are officially mandated.

As an important starting point, "human subjects research" 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. In the U.S., the official definition of human subjects research and the rules surrounding it grew out of abuses that took place in the absence of formalized regulation when researchers convinced themselves that the benefits of their studies outweighed what should have been obvious harms.[7] At universities (and many other organizations), the proper conduct of human subjects research are overseen by a committee called an Institutional Review Board, or IRB.

[6] It's worth noting that remote treatment is problematic to do because mental health providers are required to hold a license in the state where the patient is located, plus there are additional obstacles in terms of privacy-compliant teleconferencing and additional burdens imposed by insurance companies. The COVID-19 pandemic has led to easing up on some of these things, but significant obstacles to remote assessment and treatment remain (Rebecca Resnik, personal communication).

[7] For a brief introduction, see, e.g., `https://journalofethics.ama-assn.org/article/history-and-role-institutional-review-boards-useful-tension/2009-04`.

Formally speaking, this project is actually not human subjects research, for two reasons. First, in general class assignments 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, this project in particular doesn't involve human subjects research according to the formal definition: we are working with publicly available social media behavior, which involves neither intervention, nor interaction with individuals, nor individually identifiable information, since Reddit is an anonymous social media site and the data have gone through another layer of automatic de-identification as an additional safeguard.

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.

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 — albeit most likely no greater or less than what you might encounter in ordinary daily life. Make sure you have read the big warning on the first page of this project and don't hesitate to let me know of any concerns.

If you are interested in further reading about the ethics of research on social media ask me; there are a lot of new papers emerging.

2.2 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 your proper use of the dataset. If you are unable to meet these conditions please select Project C, not Project A or B as described here. If you have decided to do Project A or Project B, please send me the statements below followed by the names of the member of your group as a signature. Please then also do that again in the final project writeup.

1. We have read Benton et al. (2017).
2. We understand that privacy of the users and their data is critical, and absolutely no attempt can be made to de-anonymize or interact in any way with users.
3. We understand that this project is being done solely for educational purposes, and the results cannot be used directly in research papers. If we get promising results and would like to develop the ideas into a research paper for publication, or to use what we have done further for another class, we will talk with Prof. Resnik about obtaining suitable Institutional Review Board review. (It's not hard.)
4. We understand that we may not use these data for any purpose other than this specific class project. We will not show or share this data with anyone outside class, nor do any research or development on this dataset outside the scope of the class project. If there are things we are interested in doing with this dataset outside the scope of the class project, we will talk with Prof. Resnik.
5. We will store the dataset and any derivatives on computers that require password access. If we are working in an environment where other people can log in, e.g. a department server, we will set file permissions restrictively so that only you have access. You can also use group permissions limited

to members of your group — but under no circumstances will data related to this project be world-readable.

6. Any copies of the data or derivatives of it will be accompanied by a clear README.txt file identifying Prof. Resnik as the contact person and stating further re-distribution is not to take place without contacting him first. If anyone we know is interested in the dataset, we will refer them to Prof. Resnik, rather than providing the data ourselves.

7. Once we have completed the project, we will delete any copy of the dataset we have made, including any derived files (e.g. tokenized versions of the documents).

8. We *will not* cut/paste any text content from this dataset into our project proposal, project writeup, onto the class discussion board, into e-mail, etc. If we want to identify a specific posting, e.g. in discussion on the class discussion board, we will use the ID from the dataset. If we want to give examples, we will create a paraphrase instead of the original text. For example, if a posting said *What's this world come to? http://t.co/ XxI4QnMew* we 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 your final project writeup please also include the following statement: *We have deleted all our copies of the project dataset.*

And again, if you have any questions or concerns, of course please speak with me.

3 Background on the problem

3.1 Some prior computational work

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 at and I'm happy to answer questions; please feel free to also use the class discussion board to talk about this.

Shing et al. (2018) introduce the dataset we are using in this project (see Section 3.3, below). That paper also does some modeling but that aspect of the paper is not really worth looking at. Instead, see Zirikly et al. (2019) for a description of a community-wide shared task using parts of the data — that overview talks about the kinds of approaches people took, and you can follow the references there to look at particular papers. Note that in that shared task, the teams were evaluated only on moderate-quality crowdsourced risk assessments. That might be interesting for purposes of direct comparison with the prior work, but for this project, especially the summative evaluation of predictive modeling, it will be more interesting and relevant for you to use the data with high quality expert judgments.

As some general background, 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 O'Connor (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), and are substantially similar to what you're looking at in this project.

As one nice example of exploratory data analysis specifically connected with suicidality, see Coppersmith et al. (2016); see also the Venn Clouds visualization, which is interesting `https://github.com/coppersmith/`

vennclouds). Andy Schwartz (https://scholar.google.com/citations?hl=en&user=Na16PsUAAAAJ&view_op=list_works&sortby=pubdate) has also done some nice work connected with (relatively shallow) language analysis in social science and particularly mental health, including a number of different publications and a toolkit that might be worth a look (https://www.aclweb.org/anthology/D17-2010.pdf). De Choudhury et al. (2016) look at signals that indicate that a user on Reddit mental health forum is likely to later post on SuicideWatch.

For predictive modeling specifically connected with suicide risk, one of the best papers I've seen is Coppersmith et al. (2018), though note that they are working with actual outcomes data (whether suicide was attempted or not), as opposed to risk ratings. As other pieces of recent work also may be 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, and the paper is a great example to follow of high quality work in this domain that obtained strong results. Vioulès et al. (2018) applied a similar data collection approach to the one we took with Reddit, in their case searching Twitter for 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 prediction. De Choudhury et al. (2016) are interesting from a predictive modeling perspective as well, particularly because they look at a population of people who are already discussing mental health issues and then escalate to posting on SuicideWatch.

3.2 Background on 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. 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). 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., 2015).

Note, however, that 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. Informed by discussion with several experts in assessment of suicidality, here is one grouping of relevant factors that generally 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).

- 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
  - A sense of agitation, not being able to "stand still" physically or mentally Popovic et al. (See also 2015)
  - Indications of being impulsive; risky behavior (e.g. reckless driving, promiscuity)
  - Expressing lack of hope for things to get better
- 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
  - 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

Vioulès et al. (2018) have a related discussion of risk factors and warning signs.

In addition, there is some quite recent work that has been attempting to formalize and validate suicidal crisis in diagnostic terms as a mental state that is specifically characteristic of imminent suicide risk, as distinguished from the mental state associated with depression or lifelong suicide risk. A particularly promising line of work involves a definition of *suicide crisis syndrome*. To be identified as having suicide crisis syndrome, the individual must meet both criterion A and two of the criteria from B:

- Criterion A: Frantic hopelessness or state of entrapment defined as being stuck in a life situation that is painful and intolerable, and a feeling that all routes of escape are blocked.
- Criterion B:
  - Affective dyscontrol, including emotional pain or mental pain; severe panic with agitation, and dissociation; rapid mood swings that can include happiness; and acute anhedonia.
  - Cognitive dyscontrol, which can include ruminative flooding associated with headache or head pressure; cognitive rigidity; and inability to suppress the ruminative thoughts. (For example, you might assess by asking: "Do you control the thoughts or do the thoughts control you?")
  - Overarousal with insomnia and agitation.
  - Social withdrawal and isolation, and evading communication.

See `https://www.mdedge.com/podcasts/psychcast/dr-igor-galynker-identifying-suicide-crisis-syndrome-part-1` for discussion and some references.

3.3 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'. Shing et al. (2018) provides a detailed description of how the dataset was constructed from a large collection of every publicly available Reddit posting from January 1, 2008 through August 31, 2015 (with partial data from 2006-2007). Briefly:

- We identified the 11,129 users who had ever posted to r/SuicideWatch (which we'll sometimes abbreviate as SW), a discussion forum where the aim is to provide peer support for people considering suicide. Posters on SW generally tend to fall into one of three categories': people who themselves are considering the possibility of self-harm, people who are worried about a friend or loved one, and people who want to help. The fact that someone posted to SuicideWatch can be viewed as a form of indirect supervision, i.e. a noisy indicator of possible suicidality.
- Of those 11,129 users, a subset of 934 were randomly selected, and crowdsource workers labeled each individual on a four-point scale for risk based on reading their SW postings. Those risk labels can be viewed as a moderately reliable labeling for risk, based on our analysis of inter-rater agreement.
- Of those 934 users, a subset of 242 were labeled on the four-point risk scale by four experts in suicide prevention looking at their SW posts. Each individual was looked at by all four experts, and the inter-expert reliability (agreement on risk levels) was high, so their consensus ratings can reasonably be considered ground truth.
- As a "control" group, we identified 934 users who never posted on SW or on any other mental health related subreddits.

Note that for all of these users, in addition to posts on SW we also have all the posts they ever posted anywhere on Reddit.

Our four-level categorization of risk adapts Corbitt-Hall et al. (2016), who provided lay definitions based on risk categories in Joiner Jr et al. (1999). The levels are:

(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). Note that for purposes of this project, we are going to be adapting from a 4-way labeling scheme to binary classification (though you can also work with four labels and map to binary at the end if you prefer).

Also, for purposes of this project, you should exclude any posts to 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 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.

3.4 Additional data

Although it may or may not be useful for this project, we also provide a subset of data from the MyPersonality project,[8] 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.[9]

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.[10] The last of these, neuroticism, is a predictor of depression.[11]

4 The project problem

There are two alternatives to choose from: exploratory data analysis using computational linguistics methods and models, and supervised learning to distinguish severe-risk users posting on SuicideWatch from users who are not at severe risk. *Your group only needs to do Project A or Project B, not both.*

4.1 Project A: Exploratory data analysis

4.1.1 Ideas for potentially relevant features of language

The goal here is for you to go deeper than you have so far with techniques for exploring differences in language use. Are there detectable differences in the language of users who are high risk versus those who are not?

To tackle this, you should look identify features of language that you think might be worth exploring in social media for identifying 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. Background on potentially relevant features appears above on Section 3.2. 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

[8] `mypersonality.org`

[9] 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 was that data from 87 million people was improperly shared (`https://www.npr.org/sections/thetwo-way/2018/04/04/599542151/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?

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

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

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 1: LDA-induced themes related to depression.

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.

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.

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*).[12] Empath (Fast et al., 2016) is another interesting alternative to consider; in addition using pre-built categories, it supports building your own categories using a set of seed terms.[13]

Topic models. Topic models provide one way to capture content-related generalizations that might be valuable in helping to characterize high-risk users. As a related example, 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 1 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. There are a variety of topic modeling packags out there now; two that are widely used include MALLET and Gensim. In addition, Viet-An Nguyen's Segan

[12] 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).

[13] Code: `https://github.com/Ejhfast/empath-client`; demo `http://empath.stanford.edu/`

package (`https://github.com/vietansegan/segan`) implements a number of variants of topic models with particularly clean code (in Java) and easily understandable file formats, including the ability to incorporate prior knowledge using "seeded" distributions and the version of supervised LDA that I discussed in class, in which the regression model employs both topic-level and lexical weights.

Readability measures. De Choudhury et al. (2016) 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.[14]

How might syntactic variation be related to depression or suicidality? One could use computational linguistics methods to explore, for example, 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). 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, in order to operationalize 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

[14] 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).

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.

Other non-language measures might take into account characteristics of the individual available on Reddit. For example, one possibility would be to estimate personality characteristics from their history of posts, as done by Mairesse et al. (2007) and others since then (see discussion of the MyPersonality dataset in Section 3.3), and then use those estimates in the classifier. Another possibility could be to use profiles of people's posting behavior as an indication of individual characteristics, e.g. someone who posts frequently to Reddit groups involving firearms, military service, and weightlifting may manifest suicidal intent using language very differently from someone else who posts on groups discussing college admissions, relationships, and anorexia.

Visualizations. There are many interesting ways to visualize relevant information — feel free to explore some of these, though please also resist the temptation to get sucked into visualization itself too much and not spend enough time thinking about the problem itself from an NLP perspective. One visualization that's been used in the mental health setting that's easy and interesting is "Venn clouds"; see `https://github.com/coppersmith/vennclouds`. (It's also worth noting that although a lot of people really dislike word clouds as a visualization method for text, in practice I've found that one really practical way to use them is as a visualization for topic models, where each topic gets its own cloud and the "weight" of a word in a topic is used in place of frequency.)

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). And of course 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.

4.1.2 Defining contrasts for analysis

You have some flexibility in how you decide to draw contrasts for purposes of your analysis.

Groups of individuals. One thing to think about is, which groups of individuals are you aiming to contrast? One possibility would be to limit your exploration to the users labeled by experts, whicih helps for reliability but reduces the amount of data available. Another possibility would be to look at severe versus non-severe risk in the larger number of crowdsource-labeled individuals – or perhaps to begin there, formulate tentative conclusions or insights, and then see to what extent those hold up when you look at the expert-labeled data. Or, related, you could look specifically at differences in the labeling by crowdsourcers versus experts, to get insight into how the two differed; we found in Shing et al. (2018), for example, that crowdsourcers tended to err on the side of higher severity, but we didn't look at properties of the posts that might have led them to do so.

Note that if this were a project that involved exploratory data analysis *and* predictive modeling, then using the test data for exploratory analysis would be prohibited. For purposes of this class project, however, if you are doing exploratory work you are welcome to use the expert-labeled individuals. It will just be important to make sure that no insights from that ever go into developing predictive models that are evaluated on the same dataset.

4.2 Project B: Supervised classification

The goal here is to develop a classifier to identify "positive" users from controls, using linguistic and possibly other features.

4.2.1 Defining the classification task

The shared task conducted by Zirikly et al. (2019) used *only* the crowdsourced data, but for this project you should use the 242 expert-labeled individuals as your test set. I would recommend that you hold back the test set for your summative evaluation, using the crowdsource-labeled data to develop your classification approach and perform formative evaluations. Note that if you do this, a useful choice to make would be to use the crowdsourcers-only train/test split from Zirikly et al. (2019), which would enable you to compare your performance with the numbers reported by participants in that shared task.

Even with that, there are actually several different possible tasks you could be doing, depending what information you make availabe to the system. Zirikly et al. define three variations:

- **Task A** is about risk assessment: the task simulates a scenario in which there is already online evidence that a person might be in need of help, and the goal is to assess the degree of risk from what they posted. This task uses the smallest amount of data, with each user typically having no more than a few SuicideWatch posts. This would be a binary classification of severe-risk (d) versus the lower-risk categories (a-c).

  This task approximates real-world situations on moderated peer-support groups, where you want someone who is at imminent risk to be drawn to the attention of a moderator as quickly as possible. Milne et al. (2019) are a good example: they assign a four-way level of urgency for moderator attention based on posts in an online peer support forum, a use case that was the focus of two shared tasks similar to this that were held in conjunction with the Workshop on Computational Linguistics and Clinical Psychology Milne et al. (2016). Although automatic classification achieved high accuracy, the most striking practical result was a dramatic improvement in moderator response time: with moderator training, introduction of automated triage cut median moderator response time significantly, and also substantially reduced response time variability.

- **Task B** is the same risk assessment problem as task A, but in addition to the SuicideWatch posts (which identify that they may need help), you can also use the user's posts elsewhere on Reddit, which might tell you more about them or their mental state. On average each user we collected data for has

more than 130 posts on Reddit, and the subreddit categories are wildly diverse, from *Accounting* to *mylittlepony* to *SkincareAddiction* to *zombies*. This wouldalso be a binary classification of severe-risk (d) versus the lower-risk categories (a-c).

This task approximates real-world situations where there is evidence someone might be at risk, such as a flagged posting, but also access to a more general history of their posts. It is inspired in part by Facebook's suicide prevention infrastructure (de Andrade et al., 2018), where if a post is flagged as indicating possible risk (either algorithmically or by a person), human review takes place to decide if and how intervention should take place. Gomes et al. describe a process that only involves reviewing the flagged post, but in principle a reviewer could make a more informed judgment by looking at the broader context including their recent posts. In addition, this version of the task creates the possibility of improving automated assessment by taking into account characteristics of the individual, e.g. one possibility would be to estimate personality characteristics from their history of posts, as done by Mairesse et al. (2007) and others since then (see discussion of the MyPersonality dataset in Section 3.3), and then use those estimates in the classifier; another possibility could be to use profiles of people's posting behavior as an indication of individual characteristics, e.g. someone who posts frequently to Reddit groups involving firearms, military service, and weightlifting may manifest suicidal intent using language very differently from someone else who posts on groups discussing college admissions, relationships, and anorexia.

- **Task C** is about screening. Here predictions are made only from users' posts that are *not* on SuicideWatch (even though ground truth about the individual was determined by looking at their SW posts). This task simulates a scenario in which someone has opted in to having their social media monitored (e.g., a new mother at risk for postpartum depression, a veteran returning from a deployment, a patient whose therapist has suggested it), and the goal is to identify whether they are at risk even if they have not explicitly presented with a problem. This task would be a binary classification of severe-risk (d) versus controls, and you could also train or develop your system using labeled data in categories (a-c), or not, as you see fit. (Note that this definition is different from the Zirikly et al. Task C, which used d versus a-c just as in the other tasks.)

  A real (albeit arguably misguided) example of this approach was Samaritan's Radar, a social listening application designed to scan a user's Twitter posts for key words or phrases indicating need for crisis care (Lee, 2014). The application did not allow for participant consent or opt-in; rather, it permitted account holders to monitor friends, family, and others they followed on the site without their consent, alerting the account holder to tweets of concern for later followup at their discretion. The project was shut down nine days after launch due to broad public protest, including a `changeg`.org petition. As Horvitz and Mulligan (2015) observe the goals of this nonprofit were laudable, but people viewed them as "playing fast and loose with the privacy and mental health of those it was seeking to save", leading to public backlash.

You are welcome to choose which of these three tasks you are going to work on. Task A most closely resembles standard supervised classification tasks, although in this case you might have multiple documents per individual and it is the individual, not the document, that gets labeled. Task B is a variation where the really interesting question is how you might be able to exploit more information about the individual if you had it. Task C is undoubtedly the most challenging, since the idea is to find people who are at severe risk *without* any evidence that they have reached out for help on a peer-support group.

4.2.2 Classification approach

All of the possibilities in Section 4.1 are certainly fair game in terms of features, as are the symptoms or characteristics in Section 3.2, and of course you can propose other elements of analysis that might be predictive, if you like. But the thing I definitely do *not* want you to do is to simply treat this as a generic

supervised classification problem, e.g. by dumping bag-of-words features into a large-margin classifier (the old fashioned approach) or by simply dumping the data into a standard deep learning pipeline where you start with BERT and a typical classifier architecture, fine-tune on the task, train, and test. It's fine to do things like that as *baselines* for comparison, and certainly fine to *build on* standard approaches and packages — don't reinvent the wheel! — but if you're not doing anything that involves *thinking about this particular problem and data*, you're not really doing the project that I'm assigning. I care much more about you connecting what you're doing here with what you've learned over the semester, than I do about your system actually performing well on a classification task.

As a detail to consider that's particular to this project and different from most other supervised clsasification problems, 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... Yet another possibility, which Han has pursued in his research, is to use hierarchical attention networks (Yang et al., 2016), which can aggregate hierarchically from words to sentences to documents to individuals who are labeled at the top level.[15]

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.

4.2.3 Evaluation

Evaluating classifier performance. 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. Classification should be evaluated using precision, recall, and F1-measure on held-out test data. Note that precision (and therefore also F-measure) is sensitive to the distribution of data in the test set, which means that performance on an artificial test set may not give you a good indication of how your technique would perform in the real world. (Positives are over-represented in the expert test set compared to the real world –

[15] Han's just had a paper using that approach accepted to ACL2020 and he would probably share a pre-publication version with you if you ask him nicely.

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.

Error analysis. I don't expect you to spend as much time digging around in the data as if you were doing Project A (exploratory data analysis). However, in a supervised classification setting one really good way to get into the data is via error analysis. You can do this in formative evaluations, in the process of improving your system, or you can do it as part of the summative evaluation of the system, looking at the performance of the classifier on the real test data at the end, or you can do both.

There are a variety of things you can do in error analysis. One typical approach is to build a confusion matrix, looking at the counts of errors where the true label is x and the assigned label was y. Often it is possible to look at the subset of items contributing to one cell of that table, and identify some useful generalizations. As one example, previous error analysis with the four-way classification has shown us that when people with label a (no risk) are misclassified as d (severe risk), it can be because they include a lot of language indicating high risk (e.g. *kill*, *depressed*, *reason to live*) but are talking about a friend or family, not themselves (e.g. "I'm worried about my sister killing herself. She is really depressed and keeps saying she has no reason to live."). In looking at things you want a classifier to pay attention to, e.g. symptom categories in Section 3.2, you might consider the possibility of a feature or even classification sub-task specifically devoted to identifying whether the author is talking about themselves or someone else.

5 What you need to do

Follow the top-level guidance for all projects in terms of proposal, deliverables, etc. Here are some notes worth bearing in mind particularly for Project Ideas A and B.

Project proposal. This project is deliberately underspecified: the first part of your job, assuming you've figured out who you'll work with, is to scope out a project that will be feasible within the necessary time frame. The biggest risk here is carving off a project that is too ambitious to be done with the time you have, so try to propose a project plan that describes what your group plans to try in enough detail that we can steer you away from approaches that are likely to get you bogged down. 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.

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.)

Note that you are more than welcome to use/adapt off-the-shelf code rather than implementing things yourselves. In fact, *this is strongly encouraged.* It's better to spend your time *exercising* what you've learned, not creating your own from-scratch implementations of SVM classification, LDA, deep learning classifiers, syntactic parsing, etc. 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 at the end if they were helpful to your group, saying explicitly what they did that helped.

Project writeup. See the main top-level project page for what's expected in writeups. Here's some additional guidance particularly for these project choices.

- Introduction. 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.

- Data and methods. Some recommended things to include:

  1. 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.
  2. Basic information about preprocessing, e.g. how you did tokenization, removal of stopwords (if you did that), etc.
  3. 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, modified to preserve anonymity, along with the posterior distribution of topics for each example document."

- 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.

- Evaluation

  1. 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.
  2. 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., plus discussion of how you did error analysis and what you found. 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."

3. 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. Then make sure to include the statement and typed signatures as discussed in 2.2.

- Discussion and future work

1. 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?
2. Optionally, any particular difficulties or hurdles you encountered. Please feel free to include ways in which final projects like this could be made better.

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.

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 O'Connor. 2016. Social media, big data, and mental health: current advances and ethical implications. *Current opinion in psychology* 9:77–82.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. 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, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights* 10:1178222618792860.

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 Linguistics and Clinical Psychology*. Association for Computational Linguistics, San Diego, CA, USA, pages 106–117. http://www.aclweb.org/anthology/W16-0311.

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.

Norberto Nuno Gomes de Andrade, Dave Pawson, Dan Muriello, Lizzy Donahue, and Jennifer Guadagno. 2018. Ethics and artificial intelligence: suicide prevention on facebook. *Philosophy & Technology* 31(4):669–684.

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.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. pages 4647–4657.

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 depression—theoretical 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.

Eric Horvitz and Deirdre Mulligan. 2015. Data, privacy, and the greater good. *Science* 349(6245):253–255.

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.

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.

Naomi Lee. 2014. Trouble on the radar. *Lancet* 384(9958):1917.

François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res.* 30:457–500.

David N Milne, Kathryn L McCabe, and Rafael A Calvo. 2019. Improving moderator responsiveness in online peer support through automated triage. *Journal of medical Internet research* 21(4):e11410.

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).

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, 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).

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. Association for Computational Linguistics, pages 25–36.

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

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.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1480–1489.

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.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. pages 24–33.