Computational Linguistics II Final Exam Project: Exploring Linguistic Signal for Schizophrenia

1 The problem setting

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 clinical psychology 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 [2]. 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 [1, 29]. 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]

Although automated language analysis connected with clinical conditions goes back at least as far as the 1990s [20], it has not been a major focus for computational linguistics compared with other application domains. However, there has been noticable uptick in research activity on this topic, which is consistent with, and gains power from, the recent rise in computational linguistics activity connected with computational social science more broadly. For example, one recent shared task brings together research on the Big-5 personality traits [4, 14], and another involved research on identification of emotion in suicide notes [21]. Research has been done on language analysis in the context of autistic spectrum disorders [17, 22, 31], dementia [16, 19, 27], depression [18], post-partum depression [9], general life satisfaction [28], and suicide risk [11]. Capitalizing on this rising interest, Resnik et al. organized the first ACL Workshop on Computational Linguistics and Clinical Psychology [26, CLPsych], which has become an annual event. The second CLPsych workshop

[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

in June 2015 included a shared task looking at Twitter to identify people who might have depression or post-traumatic stress disorder (PTSD) [7], and the upcoming CLPsych 2016 workshop includes a shared task involving analysis of postings on young adult discussion forums to identify authors who might require intervention by site moderators, e.g. because they pose a risk to themselves or others.

This final exam project 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 Linguistic manifestations of schizophrenia

Schizophrenia is a severe form of psychosis, a mental condition involving loss of contact with reality, and at a high level it can be described as "chronic and severe mental disorder that affects how a person thinks, feels, and behaves."[6] Its characteristics include (as a partial list): "positive" symptoms, including hallucinations, delusions (e.g. paranoia), and thought disorders, "negative" symptoms, such as flat affect (reduced expression of emotion), reduced feelings of pleasure in everyday life, and difficulty beginning or sustaining activities, and cognitive symptoms such as poor executive function (the ability to understand information and use it to make decisions), trouble focusing, and problems with working memory. A short video at `https://www.youtube.com/watch?v=bWaFqw8XnpA` shows interviews with four schizophrenia patients and identifies some of their symptoms. Before reading further, I recommend that you watch it, and I also recommend looking at the Wikipedia page on schizophrenia at `https://en.wikipedia.org/wiki/Schizophrenia`.

Abnormalities of language are a central clinical manifestation of schizophrenia. Some of these abnormalities involve surface form in language; for example, "clanging" or "clang associations" describe a mode of speech in which people might compulsively use nonsensical rhyming or alliteration, e.g. "The train brain rained on me". Others involve "thought disorder", which involves failure to maintain a coherent discourse plan or logical connections between thoughts, "word salad" or nonsensical speech, and other kinds of bizarre content. Negative symptoms are sometimes reflected in "paucity of thought", where the language doesn't contain a great deal of real content. Covington et al. [8] survey language in schizophrenia across a full range of linguistic levels from phonetics through phonology, morphology, syntax, semantics, and pragmatics.[7]

Kuperberg and Kaplan [15] provide another in-depth survey of language in schizophrenia with a particular emphasis on thought disorder, and Bedi et al. [3] provide an interesting and promising computational approach to predicting psychosis, using sequences of vector space representations to measure semantic coherence (or lack of semantic coherence) in order to identify young people who are likely to experience psychosis.

The goal of this project is to use computational linguistics techniques in exploring linguistic signal related to schizophrenia by looking at language use in social media. We'll look at this from the perspectives of both data exploration and prediction.

Note that, as defined, we are making an assumption that may or not be true, namely that linguistic manifestations of schizophrenia are actually visible in social media. Some linguistic reflexes of schizophrenia are clearly not available, e.g. flattened intonation is a negative symptom of schizophrenia that would not be visible in written language. Others might be present but difficult to identify by looking at people's individual social media postings, e.g. sentences that could make sense in isolation but are incoherent or inappropriate in the context of the discourse or conversation with other people. It's also possible that in written language

[6] http://www.nimh.nih.gov/health/topics/schizophrenia/

[7] Full PDF can be found at `http://ai1.ai.uga.edu/caspr/litreviewSR-published.pdf`

in general, where people have the chance to edit, may show less evidence of schizophrenic symptoms than spontaneous speech.

On the other hand, a review of the literature suggests many linguistic correlates of schizophrenia that could potentially show up in social media language use. One example might be higher-than-normal use of neologisms (newly made-up words or expressions). Another is a reduction in syntactic complexity, e.g. phrasal complexity as measured by looking at use of clausal conjunctions and clause embedding. And it's also possible that social media postings by schizophrenics might show evidence of the lack of semantic coherence explored by Bedi et al. [3] (see also papers cited by Bedi et al.).

3 Datasets

Even when anonymous, postings on social media can be deeply personal and potentially upsetting, particularly where mental health issues are involved, and data needs to be handled with great sensitivity. It is absolutely essential that you read and understand Section 7, below.

3.1 Qntfy schizophrenia dataset

The primary dataset for this project is a set of nearly 1M tweets collected by Glen Coppersmith, founder of Qntfy.io, a company bringing large-scale data analysis to problems in mental health. The dataset contains data from 137 users who self-reported a relevant diagnosis on Twitter, e.g. tweeting "My doctor diagnosed me with schizophrenia", using pattern matching to capture a range of variations in how this might have been self-reported. For convenience I will refer to these as the "schizophrenia users", although it is important to note that this kind of Twitter self-report does not necessarily mean the user has had a clinical diagnosis of schizophrenia, or, even if they did, that the diagnosis is current at the time that people are tweeting. The dataset also contains data from 137 controls, people who were matched one-to-one with the schizophrenia users by age and gender (based on automatic prediction) but whose available Twitter history does not self-report a schizophrenia diagnosis. Since schizophrenia is a fairly rare (around 1% of the population), it is reasonable to assume that members of the control set are categorized appropriately.

The Qntfy dataset has a master CSV file identifying users in the schizophrenia and control groups by anonymous ID, along with gender and age estimates. It also assigns each user to one of ten folds for cross-fold validation — this assignment is nice because it means that all groups within the class can do cross-validation the same way, enabling comparison of results.

3.2 Other Twitter data

We can provide other Twitter data in order to provide a much larger baseline sample of Twitter language.

3.3 MyPersonality

Although it may or may not be useful for this project, we are also providing 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, [13]), or depression.

[8] `mypersonality.org`

People filled out the questionnaires as a fun 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]

Unfortunately the clinical measurements in the dataset do not include measures directly related to schizophrenia. However, it might be worth taking a look at what's in the dataset to see if there is information that might be useful.

3.4 Reddit

Reddit is an anonymous social media site (`http://en.wikipedia.org/wiki/Anonymous_social_media`) organized essentially as an online bulletin board system. It contains discussion categories or "subreddits" on specific topics, e.g. gaming, food, personal finance. The format for subreddit naming is */r/NAME*. We can provide python code that uses the Reddit API (`https://www.reddit.com/dev/api`) to collect Reddit postings, e.g. if you want to explore whether there is useful information on the subreddit /r/Schizophrenia (although, of course, there is no guarantee that a user posting to this subreddit is actually suffering schizophrenia). We can also provide a set of "control" postings from other Reddit forums.

4 The project problem

As noted in Section 1, algorithms that identify signals of depression from individuals' informal language could have a significant impact. There are two components to this project: exploratory data analysis using computational linguistics methods and models, and supervised learning to distinguish depressed from non-depressed users.

4.1 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. For example, are there detectable differences in the language of the "schizophrenia users" vs. controls in the Qntfy dataset?

To tackle this, you should look at papers cited in Section 2, identify features of language that you think might be worth exploring in social media as potential differentiators of schizophrenia from control 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.

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. One could also look at distinctions involving neologisms, e.g. by establishing a large baseline vocabulary (using dictionaries and large baseline datasets) and seeing how often people use words outside that vocabulary. It's possible that n-gram language models could potentially pick up differences in use compared to typical language use.

[9]The subset of MyPersonality data is being used in this project with permission of the researchers who created the dataset.

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.

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, [20]) makes it possible to look at word categories like NEGEMO (negative emotion words) or INSIGHT (including words like *accept*, *admit*, *believe*, *conclusion*, *explanation*). Fineberg et al. [10] use LIWC to explore differences in word class use associated with schizophrenia; see also [12]. It's also possible that WordNet categories could be useful here.[10]

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. [24] looked at emotional instability and depression using topic models in a corpus of writing by college students [20]. 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 might be worth considering.

Table 2 illustrates topics obtained by running a 50-topic *supervised* topic model (sLDA) on the Pennebaker stream-of-consciousness dataset [23]. 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 [13]. 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 2 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 ($p < .005$).

Other forms of dimensionality reduction. Bedi et al. [3] use latent semantic analysis (LSA) as a way to capture semantic content, as a way of operationalizing the idea that people suffering schizophrenia often

[10] `https://wordnet.princeton.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 2: 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.

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 semantic trajectories within or across tweets — one obvious thing to explore is whether the cross-clause measures that Bedi et al. used in their study could be applicable here.

Syntactic measures. As noted above, syntactic simplification is one factor that has been associated with schizophrenia. Twitter is harder to parse than many other forms of language, but it could be worth exploring whether either general tools (e.g. Stanford CoreNLP) or Twitter-specific tools (e.g.Tweet NLP, `http://www.cs.cmu.edu/~ark/TweetNLP/`) provide sufficient quality to measure elements of syntactic complexity.

Readability measures. Some standard measures of "readability" capture lexical and/or syntactic factors. See, e.g., `https://pypi.python.org/pypi/readability/0.1`.

Non-language measures. It would be perfectly reasonable to explore non-language characteristics such as age and gender. Other characteristics to look at could include 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).

These are just a few ideas — you should look at papers related to schizophrenia 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 schizophrenic 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 comonents analysis to take a larger set of features and reduce it to 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.

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 that in principle, exploratory data analysis should exclude test data, to avoid coming up with ideas for features that are overfitted to the test set. However, for this project, the dataset of positive instances is really small, so there's really no choice but to look at the available data and use cross-validation for testing (see below).

4.2 Supervised classification

The second goal of this project is to explore a supervised learning approach to distinguishing "schizophrenic users" from controls, using linguistic and other features. Classification should be evaluated using precision, recall, and F-measure. Optionally, it is also not uncommon in this area to generate receiver-operator characteristic (ROC) curves and use AUC (area under the curve) as an evaluation metric, http://en.wikipedia.org/wiki/Receiver_operating_characteristic. For some nice examples of published research of this kind, see [6, 5].

All of the possibilities in Section 4.1 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 [25] when it comes to supervised learning methodology, including, for example, evaluating improvements against lower-bound baseline (such as unigram features), using cross-validation to keep training and test data separate, etc.

5 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. Following the outline of the project writeup (below) would be one good way to organize this.

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, even though we have too little data for a proper train/test split. (See note at the end of Section 4.1.)

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*. If you prefer to provide a link to a code repository that's fine, please just remember that the datasets themselves 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 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 on computational linguistics for mental health, e.g. by Coppersmith et al. and De Choudhury et al. [6, 5], 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. Please also provide a rough breakdown of people's roles in the project. If roles weren't broken out cleanly, that's ok, I'm mainly just interested in how you organized things.
      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.

   (b) Data and Methods
      i. The data and resources you used. How you got it, basic properties (size, etc.), and, if applicable, anything you needed to do with data you received 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. Cite relevant sources, e.g. Kuperberg and Caplan [15], as appropriate. 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. For example, 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., `http://cran.r-project.org/web/packages/lda`)) with 40 topics, using each author's combined set of tweets as the document, and that author's group (-1 for schizophrenic, +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 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 one."

(c) Relevant information about any other algorithms and models, e.g. PCA, supervised classification, etc. Identify what approach you too 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, how you evaluated what you did in development and final testing, e.g. including details of cross-validation, evaluation metrics, etc. For discussion of evaluation metrics and presentation of results, see Lin and Resnik *Evaluation in NLP*.

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. [3]."

(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 exam 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    rating1    rating2    rating3    Justification
your_name    teammate2_name    rating1    rating2    rating3    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.

6 Grading

The group will receive a grade-in-common out of 75 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 25 points, each team member should anonymously rate each other team member as follows – *making sure to look at the definitions below to see what the numeric scales are supposed to mean and how to calibrate them.* Let me be clear: I'm very serious about the calibration issue. I want to see detailed justifications for anything other than the average rating. If everyone simply gives everyone a high score on every criterion, or if you provide insufficient justification for your ratings, I will either kick this back to you and make you do it again, or I will simply replace the number you gave me with the number that I believe appropriate based on the evidence you've given.

- Rating 1: Collaboration. Scale from 1 to 10 where 10 is highest.
- Rating 2: Contribution to success of the project. Scale from 1 to 10 where 10 is highest.
- Rating 3: Effort. Scale from 1 to 5 where 5 is highest.

Collaboration. 10 means that this person was so incredible to collaborate with that you would take extraordinary measures to collaborate with them again, and 1 means you definitely would avoid collaborating with this person again. Give 5 as an *average* rating for someone who was fine as a collaborator, but for whom you wouldn't feel strongly about either seeking them out or avoiding them as a collaborator in the future. *I expect to see a detailed justification, with examples, for any number higher than 5.*

Contribution. 10 means that this person did their part and more, far over and above the call of duty. 1 means that this person *really* did not contribute what they were supposed to. Give 5 as an average score showing that the person did what was expected. Note that this is a subjective rating relative to what a

person was *expected* to contribute. If five people in a group were contributors, and each did as they were expected on their pieces, then all five should get a rating of 5; you would not slice up the pie 5 ways and give each person a 1 because they each did 20% of the work! It is your job as a group to have understood what the expected contributions will be, to make sure everyone is comfortable with the relative sizes of the contributions, and to recalibrate expectations if you discover you need to. Try to keep things as equitable as possible; for example, one person's skills might mean they could do a total of 10% of the work compared to another person's 15%, and both people could get the same "they did what was expected" as long as everyone was ok with the expectations. If you run intro trouble breaking up tasks, agreeing on expectations, etc., I would be happy to help the group in working these things out. *I expect to see a detailed justification, with examples, for any number higher than 5.*

Effort. A rating of 3 should be average, with 5 as *highly unusual and superior effort* (whether or not they succeeded) and 1 as *didn't put in the effort.* A rating below 3 would not be expected if the person's contribution was 5 or better. If a person just didn't manage to contribute what they were expected to, but you think they did everything in their power to make it happen, you could still give them an average or even above-average value (the classic "got an A for effort") even while giving them a low contribution score. *I expect to see a detailed justification, with examples, for any number higher than 3.*

Justification. The Justification column should contain a statement justifying in detail any number higher or lower than the expected or average value as specified above. You should be using this sparingly. An example: *I gave Philip a 7 for collaboration because he went well beyond the minimum in helping out with other people's parts of the project. As a specific example, when I was getting completely uninterpretable results from LDA, he spent literally hours with me over several days working side by side to figure out how to optimize the hyperparameters, which turned out to make a big difference. Unfortunately, I think his own contribution may have suffered as a result: I gave him a 4 for Contribution because he ended up contributing 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 never actually got to the autoencoders. I think the higher score on collaboration offsets the lower score on contribution, though.*

7 Ethical use of data

7.1 General notes

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 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.

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 inappropriate to use Web searches to look for further information from or about a user in these datasets, even for benign purposes. If you have any questions about appropriate use of data, please let 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, please let me know.

If you are interested in further discussion about the ethics of research on social media, Solberg [30] has some interesting discussion, and so does 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).

7.2 Use of the Qntfy dataset

The primary dataset for this educational project has kindly been provided by Glen Coppersmith and Qntfy, a company with commercial efforts in mental health. 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 and involving Glen as a collaborator.
3. You may not use these data for any purpose other than this specific class project. You may 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 tweets). 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 need to identify a specific posting, use the tweet ID from the JSON, e.g. 481691063930045663 4. If you want to give examples, please create a paraphrase instead of the original text. For example, if the original tweet is *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 tweet 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 Qntfy 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 Qntfy dataset.*

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

8 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 mailing list or 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

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

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

[3] 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. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1, 2015.

[4] Fabio Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski. Computational personality recognition (shared task) workshop. In *International Conference on Weblogs and Social Media*. AAAI, 2013.

[5] Munmun De Choudhury, Scott Counts, Eric Horvitz, and Michael Gamon. Predicting depression via social media. In *AAAI*. AAAI, July 2013.

[6] Glen Coppersmith, Mark Dredze, and Craig Harman. 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, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[7] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 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*, Denver, Colorado, USA, June 2015. North American Chapter of the Association for Computational Linguistics.

[8] MA Covington, C He, C Brown, L Nai, JT McClain, BS Fjordbak, J Semple, , and J Brown. Schizophrenia and the structure of language: the linguist's view. *Schizophrenia Research*, 77(1):85–98, September 2005. doi:10.1016/j.schres.2005.01.016.

[9] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems*, pages 3267–3276. ACM, 2013.

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

[11] Derrick Harris. Darpa-funded project uses social media to predict suicide by soldiers. *GigaOM*, July 2013. http://gigaom.com/2013/07/09/darpa-funded-project-uses-social-media-to-predict-suicide-by-soldiers/.

[12] Kai Hong, Christian G Kohler, Mary E March, Amber A Parker, and Ani Nenkova. 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*, pages 37–47. Association for Computational Linguistics, 2012.

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

[14] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[15] G. Kuperberg and David Caplan. Language dysfunction in schizophrenia. *Neuropsychiatry*, pages 444–466, 2003.

[16] Maider Lehr, Emily Tucker Prud'hommeaux, Izhak Shafran, and Brian Roark. Fully automated neuropsychological assessment for detecting mild cognitive impairment. In *INTERSPEECH*, 2012.

[17] Maider Lehr, Izhak Shafran, Emily Prudhommeaux, and Brian Roark. Discriminative joint modeling of lexical variation and acoustic confusion for automated narrative retelling assessment. In *Proceedings of NAACL-HLT*, pages 211–220, 2013.

[18] Yair Neuman, Yohai Cohen, Dan Assaf, and Gabbi Kedma. Proactive screening for depression through metaphorical and automatic text analysis. *Artif. Intell. Med.*, 56(1):19–25, September 2012.

[19] Serguei Pakhomov, Dustin Chacon, Mark Wicklund, and Jeanette Gundel. Computerized assessment of syntactic complexity in alzheimers disease: a case study of iris murdochs writing. *Behavior Research Methods*, 43(1):136–144, 2011.

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

[21] John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K Bretonnel Cohen, John Hurdle, Christopher Brew, et al. Sentiment analysis of suicide notes: A shared task. *Biomedical Informatics Insights*, 5(Suppl. 1):3, 2012.

[22] Emily T Prudhommeaux, Brian Roark, Lois M Black, and Jan van Santen. Classification of atypical language in autism. *ACL HLT 2011*, page 88, 2011.

[23] Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 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)*, 2015.

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

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

[26] Philip Resnik, Rebecca Resnik, and Margaret Mitchell, editors. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Baltimore, Maryland, USA, June 2014.

[27] Brian Roark, Margaret Mitchell, J Hosom, Kristy Hollingshead, and Jeffrey Kaye. Spoken language derived measures for detecting mild cognitive impairment. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2081–2090, 2011.

[28] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Megha Agrawal, Gregory J. Park, Shrinidhi K. Lakshmikanth, Sneha Jha, Martin E. P. Seligman, and Lyle Ungar. Characterizing geographic variation in well-being using tweets. In *Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*, 2013.

[29] Kathleen Sibelius. Increasing access to mental health services, April 2013. http://www.whitehouse.gov/blog/2013/04/10/increasing-access-mental-health-services.

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

[31] Jan PH Van Santen, Emily T Prud'hommeaux, Lois M Black, and Margaret Mitchell. Computational prosodic markers for autism. *Autism*, 14(3):215–236, 2010.