Towards Developing an Annotation Scheme for Depressive Disorder Symptoms: A Preliminary Study using Twitter Data

Danielle L Mowery
Department of Biomedical Informatics
University of Utah
421 Wakara Way
Salt Lake City, UT 84115
danielle.mowery@utah.edu

Craig Bryan
Department of Psychology
University of Utah
260 S. Central Campus Dr
Salt Lake City, UT 84112
craig.bryan@utah.edu

Mike Conway
Department of Biomedical Informatics
University of Utah
421 Wakara Way
Salt Lake City, UT 84115
mike.conway@utah.edu

Abstract

Major depressive disorder is one of the most burdensome and debilitating diseases in the United States. In this pilot study, we present a new annotation scheme representing depressive symptoms and psycho-social stressors associated with major depressive disorder and report annotator agreement when applying the scheme to Twitter data.

1 Introduction

Major depressive disorder — one of the most debilitating forms of mental illness — has a lifetime prevalence of 16.2% (Kessler et al., 2003), and a 12-month prevalence of 6.6% (Kessler and Wang, 2009) in the United States. In 2010, depression was the fifth biggest contributor to the United State's disease burden, with only lung cancer, lower back pain, chronic obstructive pulmonary disease, and heart disease responsible for more poor health and disability (US Burden of Disease Collaborators, 2013).

Social media, particularly Twitter, is increasingly recognised as a valuable resource for advancing public health (Ayers et al., 2014; Dredze, 2012), in areas such as understanding population-level health behaviour (Myslín et al., 2013; Hanson et al., 2013), pharmacovigilance (Freifeld et al., 2014; Chary et al., 2013), and infectious disease surveillance (Chew and Eysenbach, 2010; Paul et al., 2014). Twitter's value in the mental health arena — the focus of this paper — is particularly marked, given that it provides access to first person accounts of user behaviour, activities, thoughts, feelings, and relationships, that may be indicative of emotional wellbeing.

The main contribution of this work is the development and testing of an annotation scheme, based on DSM-5 depression criteria (American Psychiatric Association, 2013) and depression screening instruments[1] designed to capture depressive symptoms in social media data, particularly Twitter. In future work, the annotation scheme described here will be applied to a large corpus of Twitter data and used to train and test Natural Language Processing (NLP) algorithms.

The paper is structured as follows. Section 2 describes related work. Section 3 sets out the methodology used, including a list of semantic categories related to depression and psycho-social stressors derived from the psychology literature, and a description of our annotation process and environment. Section 4 presents the results of our annotation efforts and Section 5 provides commentary on those results.

[1] For example, the 10-item HANDS scale (Harvard Department of Psychiatry/NDSD) (Baer et al., 2000).

Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 89–98,
Denver, Colorado, June 5, 2015. ©2015 Association for Computational Linguistics

2 Background

2.1 Mental Health, NLP, and Social Media

Significant research effort has been focused on developing NLP methods for identifying mental health risk factors. For example, Huang et al., in a large-scale study of electronic health records, used structured data to identify cohorts of *depressed* and *non-depressed* patients, and — based on the narrative text component of the patient record — built a regression model capable of predicting depression diagnosis one year in advance (Huang et al., 2014). Pestian et al. showed that an NLP approach based on machine learning performed better than clinicians in distinguishing between suicide notes written by suicide completers, and notes elicited from healthy volunteers (Pestian et al., 2010; Pestian et al., 2012). Using machine learning methods, Xuan et al. identified linguistic characteristics — e.g. impoverished syntax and lexical diversity — associated with dementia through an analysis of the work of three British novelists, P.D. James (no evidence of dementia), Agatha Christie (some evidence of dementia), and Iris Murdoch (diagnosed dementia) (Xuan et al., 2011).

More specifically focused on Twitter and depression, De Choudhury et al. describes the creation of a corpus crowdsourced from Twitter users with depression-indicative CES-D scores[2], then used this corpus to train a classifier, which, when used to classify geocoded Twitter data derived from 50 US states, was shown to correlate with US Centers for Disease Control (CDC) depression data (De Choudhury et al., 2013). Jashinsky et al. used a set of Twitter keywords organised around several themes (e.g. depression symptoms, drug use, suicidal ideation) and identified strong correlations between the frequency of suicide-related tweets (as identified by keywords) and state-level CDC suicide statistics (Jashinsky et al., 2014). Coppersmith et al. identified Twitter users with self-disclosed depression diagnoses ("I was diagnosed with depression") using regular expressions, and discovered that when depressed Twitter users' tweets where compared with a cohort of non-depressed Twitter users' tweets there were significant differences between the two groups in their expression of anger, use of pronouns, and frequency of negative emotions (Coppersmith et al., 2014).

2.2 Annotation Studies

Annotation scheme development and evaluation is an important subtask for some health and biomedical-related NLP applications (Conway et al., 2010; Mowery et al., 2013; Roberts et al., 2007; Vincze et al., 2008; Kim et al., 2003). Work on building annotation schemes (and corpora) for mental health signals in social media is less well developed, but pioneering work exists. For example, Homan et al. created a 4-value distress scale for rating tweets, with annotations performed by novice and expert annotators (Homan et al., 2014). To our knowledge, there exists no clinical depression annotation scheme that explicitly captures elements from common diagnostic protocols for the identification of depression symptoms in Twitter data.

3 Methods

Our first step was the iterative development of a Depressive Disorder Annotation Scheme based on widely-used diagnostic criteria (Section 3.1). We then went on to evaluate how well annotators were able to apply the schema to a small corpus of Twitter data, and assessed pairwise inter-annotator agreement across the corpus (Section 3.2).

3.1 Depressive Disorder Annotation Scheme

3.1.1 Classes

Our Depressive Disorder Annotation Scheme is hierarchally-structured and is comprised of two mutually-exclusive nodes - **No evidence of clinical depression** and **Evidence of clinical depression**. The **Evidence of clinical depression** node has two non-mutually-exclusive types, **Depression Symptom** and **Psycho-Social Stressor**, derived from our literature review (top-down modeling) and dataset (bottom-up modeling). A summary of the scheme is shown in Figure 1.

For **Depression Symptom** classes, we identified 9 of the 10 parent-level depression symptoms from five resources for evaluating depression:

[2] Center for Epidemiologic Studies Depression Scale (Radloff, 1977)

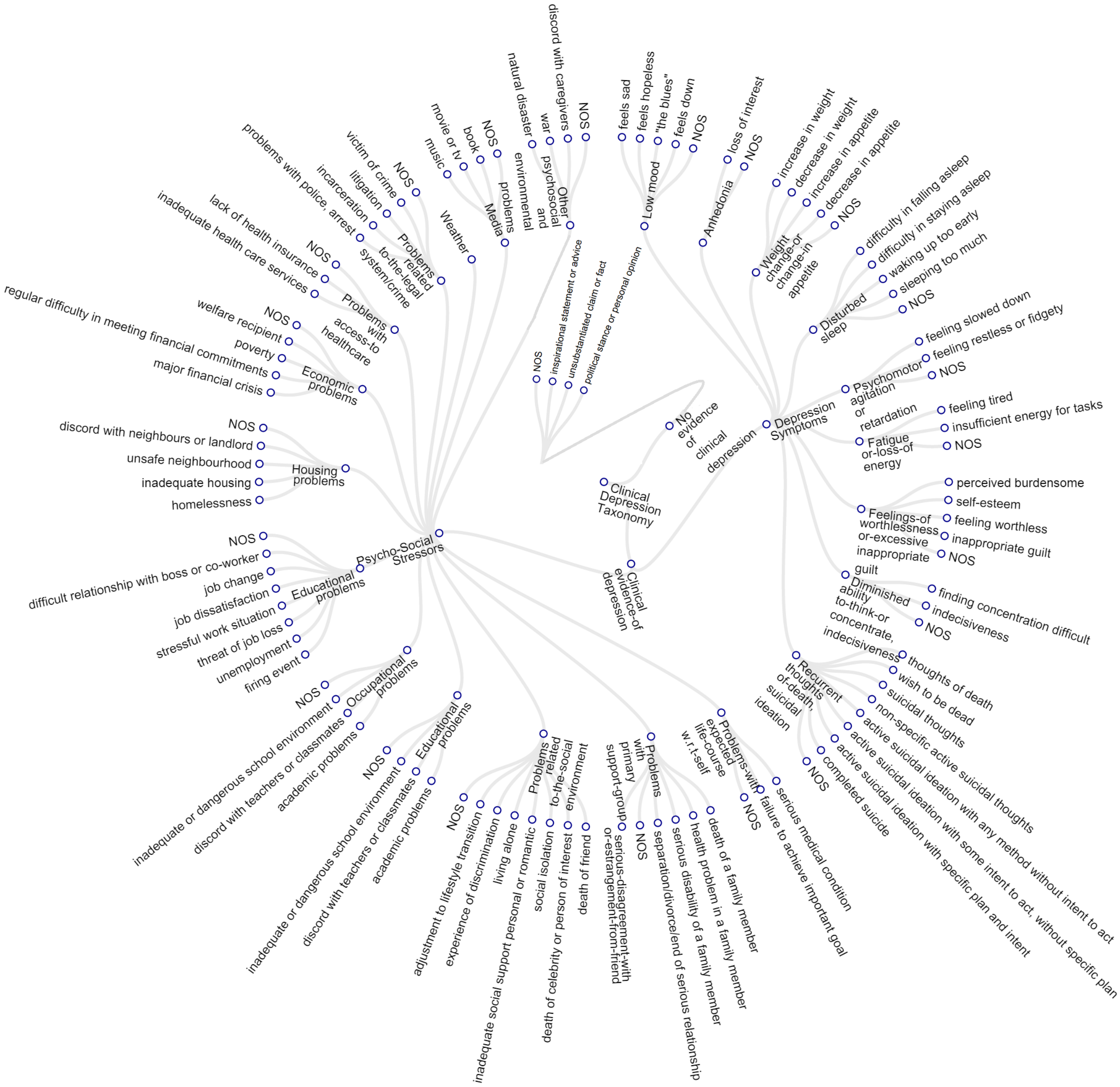

Clinical Depression Taxonomy
No evidence of clinical depression
Clinical evidence-of depression
Depression Symptoms
Low mood
feels sad
feels hopeless
"the blues"
feels down
NOS
Anhedonia
loss of interest
NOS
Weight change-or change-in appetite
increase in weight
decrease in weight
increase in appetite
decrease in appetite
NOS
Disturbed sleep
difficulty in falling asleep
difficulty in staying asleep
waking up too early
sleeping too much
NOS
Psychomotor agitation or retardation
feeling slowed down
feeling restless or fidgety
NOS
Fatigue or-loss-of energy
feeling tired
insufficient energy for tasks
NOS
Feelings-of worthlessness or-excessive inappropriate guilt
perceived burdensome
self-esteem
feeling worthless
inappropriate guilt
NOS
Diminished ability to-think-or concentrate, indecisiveness
finding concentration difficult
indecisiveness
NOS
Recurrent thoughts of-death, suicidal ideation
thoughts of death
wish to be dead
suicidal thoughts
non-specific active suicidal thoughts
active suicidal ideation with any method without intent to act
active suicidal ideation with some intent to act, without specific plan
active suicidal ideation with specific plan and intent
completed suicide
NOS
Psycho-Social Stressors
Problems-with expected life-course w.r.t-self
serious medical condition
failure to achieve important goal
NOS
Problems with primary support-group
death of a family member
health problem in a family member
serious disability of a family member
separation/divorce/end of serious relationship
serious-disagreement-with or-estrangement-from-friend
NOS
Problems related to-the-social environment
death of friend
death of celebrity or person of interest
social isolation
inadequate social support personal or romantic
living alone
experience of discrimination
adjustment to lifestyle transition
NOS
Educational problems
academic problems
discord with teachers or classmates
inadequate or dangerous school environment
NOS
Occupational problems
academic problems
discord with teachers or classmates
inadequate or dangerous school environment
NOS
Educational problems
firing event
unemployment
threat of job loss
stressful work situation
job dissatisfaction
job change
difficult relationship with boss or co-worker
NOS
Housing problems
homelessness
inadequate housing
unsafe neighbourhood
discord with neighbours or landlord
NOS
Economic problems
major financial crisis
poverty
regular difficulty in meeting financial commitments
welfare recipient
NOS
Problems with access-to healthcare
inadequate health care services
lack of health insurance
NOS
Problems related to-the-legal system/crime
problems with police, arrest
incarceration
litigation
victim of crime
NOS
Weather
Media
movie or tv
music
book
NOS
Other psychosocial and environmental problems
natural disaster
war
discord with caregivers
NOS
NOS
inspirational statement or advice
unsubstantiated claim or fact
political stance or personal opinion

Figure 1: Radial representation of Depressive Disorder Annotation Scheme

1. Diagnostic and Statistical Manual of Mental Disorders, Edition 5 (DSM-5) (American Psychiatric Association, 2013)
2. Behavioral Risk Factors Surveillance System BRFSS depression inventory (BRFSS)(Centers for Disease Control, 2014)[3]
3. The Harvard Department of Psychiatry National Depression Screening Day Scale (HANDS) (Baer et al., 2000)
4. Patient Health Questionnaire (PHQ-9) (Kroenke et al., 2001)
5. The Quick Inventory of Depressive Symptomatology (QIDS-SR) (Rush et al., 2003)

Additionally, we included a suicide related class, **Recurrent thoughts of death, suicidal ideation**, which consisted of child level classes derived from the Columbia Suicide Severity Scale (Posner et al., 2011).

For **Psycho-Social Stressor** classes, we synthe-

[3] www.webcitation.org/6Ws1k4tki

sised 12 parent-level classes based on the Diagnostic and Statistical Manual of Mental Disorders, Edition 4 (DSM IV) Axis IV "psychosocial and environmental problems" (American Psychiatric Association, 2000) and work by Gilman et al. (Gilman et al., 2013). We identified other potential parent classes based on annotation of 129 randomly-selected tweets from our corpus. The hierarchical structure of the scheme, emphasising parent and child classes assessed in this study, is depicted in Figure 2.

In the following subsections, **3.1.1.1 Depression Symptom Classes** and **3.1.1.2 Psycho-Social Stressor Classes**, we list some example tweets for each Depression Symptom and Psycho-Social Stressor class.

3.1.1.1 Depression Symptom Classes

- **No evidence of clinical depression**: political stance or personal opinion, inspirational statement or advice, unsubstantiated claim/fact, NOS
  E.g."People who eat dark chocolate are less likely to be depressed"
- **Low mood**: feels sad, feels hopeless, "the blues", feels down, NOS
  E.g. "Life will never get any better #depression"
- **Anhedonia**: loss of interest in previous interests, NOS
  E.g. "Cant seem to jam on this guitar like the old days #depressionIsReal"
- **Weight change or change in appetite**: increase in weight, decrease in weight, increase in appetite, decrease in appetite, NOS
  E.g. "At least I can now fit into my old fav jeans again #depressionWeightLossProgram"
- **Disturbed sleep**: difficulty in falling asleep, difficulty staying awake, waking up too early, sleeping too much, NOS
  E.g. "I could sleep my life away; I'm a depressed sleeping beauty"
- **Psychomotor agitation or retardation**: feeling slowed down, feeling restless or fidgety, NOS
  E.g. "I just feel like I'm talking and moving in slow motion"
- **Fatigue or loss of energy**: feeling tired, insufficient energy for tasks, NOS
  E.g. "I just cannot muster the strength to do laundry #day20 #outOfUnderwear"
- **Feelings of worthlessness or excessive inappropriate guilt**: perceived burdensome, self-esteem, feeling worthless, inappropriate guilt, NOS
  E.g. "I just can't seem to do anything right for anybody"
- **Diminished ability to think or concentrate, indecisiveness**: finding concentration difficult, indecisiveness, NOS
  E.g. "Should I do my homework or the laundry first? What does it matter anyway?"
- **Recurrent thoughts of death, suicidal ideation**: thoughts of death, wish to be dead, suicidal thoughts, non-specific active suicidal thoughts, active suicidal ideation with any method without intent to act, active suicidal ideation with some intent to act, without specific plan, active suicidal ideation with specific plan and intent, completed suicide, NOS
  E.g. "Sometimes I wish i would fall asleep and then not wake up"

3.1.1.2 Psycho-Social Stressor Classes

- **Problems with expected life course with respect to self**: serious medical condition, failure to achieve important goal, NOS
  E.g. "If it wasn't for my chronic pain, I could have made the Olympics. Now what?!"
- **Problems with primary support group**: death of a family member, health problem in a family member, serious disability of a family member, separation/divorce/end of serious relationship, serious disagreement with or estrangement from friend, NOS
  E.g. "I've been so depressed since my brother passed this year"
- **Problems related to the social environment**: death of friend, death of celebrity or person of interest, social isolation, inadequate social support personal or romantic, living alone, experience of discrimination, adjustment to lifestyle transition, NOS
  E.g. "Since Robin Williams's death, I've only known dark days"

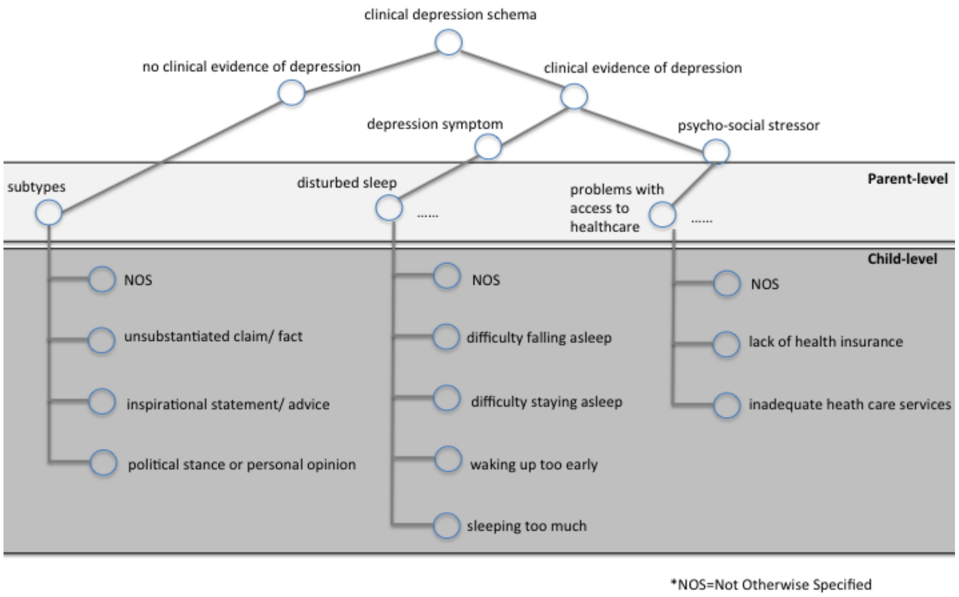

clinical depression schema
no clinical evidence of depression
clinical evidence of depression
depression symptom
psycho-social stressor
Parent-level
subtypes
disturbed sleep
......
problems with access to healthcare
......
Child-level
NOS
unsubstantiated claim/ fact
inspirational statement/ advice
political stance or personal opinion
NOS
difficulty falling asleep
difficulty staying asleep
waking up too early
sleeping too much
NOS
lack of health insurance
inadequate heath care services
*NOS=Not Otherwise Specified

Figure 2: Annotation scheme hierarchy. light gray=parent classes; dark gray=child classes. **NOS** (**N**ot **O**therwise **S**pecified indicates the parent class by default)

.

- **Educational problems**: academic problems, discord with teachers or classmates, inadequate or dangerous school environment, NOS
  E.g. “This MBA program is the worst! I feel like I’m leaving Uni with no skill sets”
- **Occupational problems**: firing event, unemployment, threat of job loss, stressful work situation, job dissatisfaction, job change, difficult relationship with boss or co-worker, NOS
  E.g. “What kind of life is this working 12 hour days in a lab??”
- **Housing problems**: homelessness, inadequate housing, unsafe neighbourhood, discord with neighbours or landlord, NOS
  E.g. “My dad threw me out of the house again. I didn’t want to live under his roof anyway”
- **Economic problems**: major financial crisis, regular difficulty in meeting financial commitments, poverty, welfare recipient, NOS
  E.g.“My clothes have more patches than original cloth. #whateverItTakes”
- **Problems with access to healthcare**: inadequate health care services, lack of health insurance, NOS
  E.g. “These generic pills do nothing to subside my depressed thoughts”
- **Problems related to the legal system/crime**: problems with police or arrest, incarceration, litigation, victim of crime, NOS
  E.g. “3 years in the joint and life hasn’t changed at all on the outside #depressingLife”
- **Other psychosocial and environmental problems**: natural disaster, war, discord with caregivers, NOS
  E.g. “I lost everything and my mind to Hurricane Katrina”
- **Weather**: NOS
  E.g. “Rainy day - even the weather agrees with my mood” [Not a DSM IV Psycho-social Stressor]
- **Media**: music, movie or tv, book, other, NOS
  E.g. “After reading Atonement I became really bummed out” [Not a DSM IV Psycho-social Stressor]

3.2 Pilot Annotation Study

The goal of this preliminary study was to assess how reliably our annotation scheme could be applied to Twitter data. To create our initial corpus, we queried the Twitter API using lexical variants of “depression" e.g., "depressed" and "depressing", and randomly sampled 150 tweets from the data set[4]. Of these 150 tweets, we filtered out 21 retweets (RT). The remaining tweets (n=129 tweets) were annotated with the annotation scheme and adjudicated with consensus review by the authors (A1, A2), both biomedical informaticists by training. Two clinical psychology student annotators (A3, A4) were trained to apply the guidelines using the extensible Human Oracle Suite of Tools (eHOST) annotation tool (South et al., 2012) (Figure 3). Following this initial training, A3 and A4 annotated the same 129 tweets as A1 and A2.

In this study, we calculated the frequency distribution of annotated classes for each annotator. In order to assess inter-annotator agreement, we compared annotator performance *between* annotators (IAA$_{ba}$ — *b*etween *a*nnotators) and *against* the adjudicated reference standard (IAA$_{ar}$ — *a*gainst the *r*eference standard) using F1-measure. Note that F1-measure, the harmonic mean of sensitivity and positive predictive value, is equivalent to positive specific agreement which can act as a surrogate for kappa in situations where the number of true negatives becomes large (Hripcsak and Rothschild, 2005). We also assessed IAA$_{ar}$ performance compared to the reference standard at both parent and child levels of the annotation scheme hierarchy (see Figure 2 for example parent/child classes). In addition to presenting IAA$_{ar}$ by annotator for each parent class, we also characterise the following distribution of disagreement types:

1. Presence/absence of clinical evidence (CE)
   e.g., **No evidence of clinical depression** *vs.* **Fatigue or loss of energy**
2. Spurious class (SC)
   e.g., false class annotation
3. Missing class (MC)
   e.g., missing class annotation
4. Other (OT)
   e.g., errors not mentioned above

4 Results

In Table 1, we report the distribution of annotated classes per tweet. The prevalence of tweets annotated with one class label ranged from 83-97%, while the prevalence of tweets annotated with two class labels ranged from 3-16%. A3 and A4 annotated all 129 tweets. Annotators annotated between 133-149 classes on the full dataset.

	A1	A2	A3	A4
1	106 (83)	116 (91)	121 (94)	125 (97)
2	20 (16)	12 (9)	8 (6)	4 (3)
3+	1 (1)	0 (0)	0 (0)	0 (0)
tws	127	128	129	129
cls	149	140	137	133

Table 1: Count (%) distribution for annotated classes per tweet; total annotated tweets (tws); total annotated classes (cls)

Table 2 shows assessed pair-wise IAA$_{ba}$ agreement between annotators. We observed moderate (A1/A2: 68; A2/A4: 43) to low (A2/A3: 30; A1/A4:38) IAA$_{ba}$ between annotators.

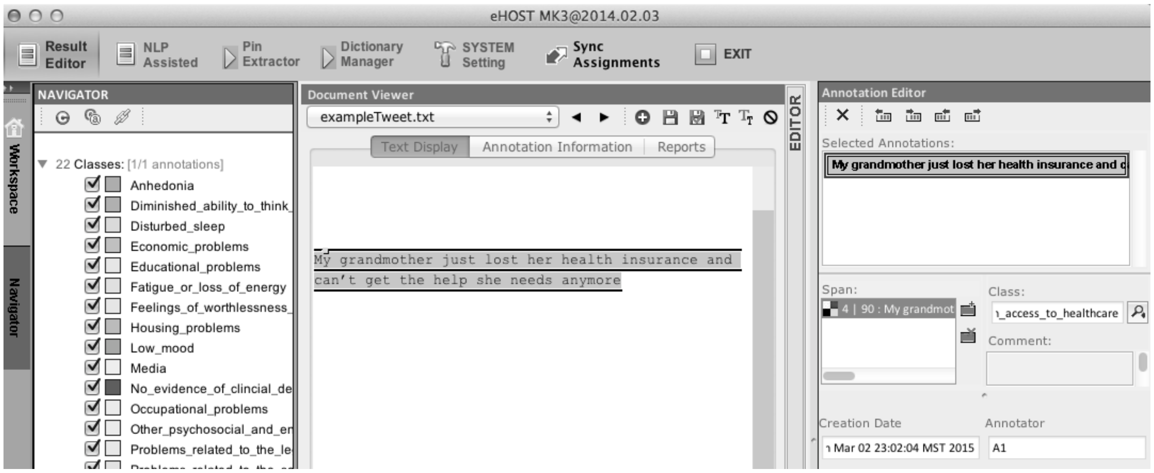
In Table 3, we report IAA$_{ar}$ for each annotator compared to the reference standard for both parent and child classes. IAA$_{ar}$ ranged from 60-90 for the parent classes (e.g. **Media**) and 41-87 for child classes (e.g. **Media: book**). The IAA$_{ar}$ difference between parent and child class performance ranged from 3-36 points.

Table 4 enumerates IAA$_{ar}$ for the observed parent classes. Note that only 12 (55%) of the parent classes were observed in the reference standard. A1 had variable agreement levels including 4 subtypes between 80-100, 6 subtypes between 60-79, and 3 subtypes between 40-59. A2 had consistently high agreement with 10 subtypes between 80-100 followed by 1 subtype IAA$_{ar}$ between 20-39 IAA$_{ar}$. A3 achieved 3 subtypes between 60-79 and 1 subtype between 40-59. A3 performed with 2 subtypes between 80-100, 3 subtypes between 60-79, 1 subtype between 40-59, and 2 subtypes between 20-39.

	A1	A2	A3	A4
A1		68	24	38
A2			30	43
A3				28
A4				

Table 2: Pairwise IAA$_{ba}$ between annotators

[4] The Twitter data analysed were harvested from the Twitter API during February 2014. Only English language tweets were retained.



exampleTweet.txt

▼ 22 Classes: [1/1 annotations]

☑ Anhedonia
☑ Diminished_ability_to_think_
☑ Disturbed_sleep
☑ Economic_problems
☑ Educational_problems
☑ Fatigue_or_loss_of_energy
☑ Feelings_of_worthlessness_
☑ Housing_problems
☑ Low_mood
☑ Media
☑ No_evidence_of_clincial_de
☑ Occupational_problems
☑ Other_psychosocial_and_er
☑ Problems_related_to_the_le

```
My grandmother just lost her health insurance and
can't get the help she needs anymore
```

Selected Annotations: My grandmother just lost her health insurance and c

Span: 4 | 90 : My grandmot

Class: ı_access_to_healthcare

Comment:

Creation Date: ı Mar 02 23:02:04 MST 2015

Annotator: A1

Figure 3: eHOST annotation tool

	A1	A2	A3	A4
parent	75	90	66	60
child	63	87	30	41

Table 3: Overall IAA_{ar} for each annotator at parent and child levels compared against the reference standard

We observed between 15-57 disagreements across annotators when compared to the reference standard (see Table 5), with **No evidence of clinical depression** accounting for 60-77% of disagreements. Missing classes accounted for 16-33% of disagreements.

5 Discussion

We developed an annotation scheme to represent depressive symptoms and psychosocial stressors associated with depressive disorder, and conducted a pilot study to assess how well the scheme could be applied to Twitter tweets. We observed that content from most tweets can be represented with one class annotation (see Table 1), an unsurprising result given the constraints on expressivity imposed by Twitter's 140 character limit. In several cases, two symptoms or social stressors are expressed within a single tweet, most often with **Low mood** and a second class (e.g. **Economic problems**).

We observed low to moderate IAA_{ba} between annotators (Table 2). Annotators A1 and A2 achieved highest agreement suggesting they have a more similar understanding of the schema than all other pair combinations. Comparing our kappa scores to related work is challenging. However, Homan et al. reports a comparable, moderate kappa (50) between two novice annotators when annotating whether a tweet represents distress.

When comparing IAA_{ar}, annotators achieved moderate to high agreement at the parent level against the reference standard (Table 3). Annotators A1 and A2 had higher parent and child level agreement than annotators A3 and A4. This may be explained by the fact that the schema was initially developed by A1 and A2. Additionally, the reference standard was adjudicated using consensus between A1 and A2. Around half of the depressive symptoms and psycho-stressors were not observed during the pilot study (e.g. **Anhedonia**, **Fatigue or loss of energy**, **Recurrent thoughts of death or suicidal ideation** — see Table 4) although may well appear in a larger annotation effort. The reference standard consists mainly of **No evidence of clinical depression** and **Low mood** classes suggesting that other depressive symptoms and psychostressors (e.g. **Psychomotor agitation or retardation**) are less often expressed or more difficult to detect without more context than is available in a single tweet. For these most prevalent subtypes, good to excellent agreement was achieved by all 4 annotators. Considerably lower agreement was observed for annotators A3 and A4 for less prevalent classes. In contrast, A1 and A2 maintained similar moderate and high agreement, respectively. In future experiments, we will leverage all annotators' annotations when generating the reference standard (i.e. the reference standard will be created using majority vote).

The most prevalent disagreement involved iden-

Parent Classes	Ct	A1	A2	A3	A4
All	148	75	90	66	60
No evidence of clinical depression	73	77	94	74	66
Low mood	52	75	91	70	63
Problems related to social environment	6	80	80	40	22
Media	4	67	33	0	31
Problems with expected life course wrt. self	3	86	0	0	0
Weather	3	86	100	0	50
Education problems	2	67	80	0	0
Disturbed sleep	1	100	100	0	100
Economic problems	1	50	100	0	0
Occupational problems	1	67	100	0	100
Problems with primary support group	1	50	100	0	0
Weight or appetite change	1	50	100	0	0
Fatigue or loss of energy	0	0	0	0	0
Housing problems	0	0	0	0	0
Psychomotor agitation or retardation	0	0	0	0	0

Table 4: Agreement for parent classes between annotator & reference standard; darker gray=higher IAA_{ar}, lighter gray=lower IAA_{ar}. Note that not all classes are listed.

	A1	A2	A3	A4
CE	25 (65)	9 (60)	36 (74)	44 (77)
MC	8 (21)	5 (33)	8 (16)	9 (16)
SC	3 (8)	1 (7)	2(4)	0 (0)
OT	2 (5)	0 (0)	3 (6)	4 (7)
Total	38	15	49	57

Table 5: Count (%) of disagreements by type for each annotator compared against the reference standard

tifying a tweet as containing **No evidence of clinical depression** (see Table 5). The line between the presence and absence of evidence for clinical depression is difficult to draw in these cases due to the use of humour ("So depressed :) #lol"), misuse or exaggerated use of the term ("I have a bad case of post concert depression"), and lack of context ("This is depressing"). In very few cases, disagreements were the result of other differences such as specificity (**Media vs Media: book**) or one-to-one mismatch (**Weather: NOS** vs **Media: book**). This result is unsurprising given that agreement tends to reduce as the number of categories become large, especially for less prevalent categories (Poesio and Vieira, 1998). We acknowledge several limitations in our pilot study, notably the small sample size and initial queried term. We will address these limitations in future work by annotating a significantly larger corpus (over 5,000 tweets) and querying the Twitter API with a more diverse list of clinician-validated keywords than was used in this pilot annotation study.

6 Conclusions

We conclude that there are considerable challenges in attempting to reliably annotate Twitter data for mental health symptoms. However, several depressive symptoms and psycho-social stressors derived from DSM-5 depression criteria and depression screening instruments can be identified in Twitter data.

Acknowledgements

We would like to thank Dr Murray Stein (University of California San Diego, Department of Psychiatry) and Dr Gary Tedeschi (California Smokers Helpline) for their comments on an early draft of the annotation scheme described in this paper. We would also like to thank Mr Soumya Smruti Mishra (University of Utah, Department of Computer Science) for help with diagram creation, and both Mr Tyler Cheney and Ms Hilary Smith (University of Utah, Department of Psychology) for their annotation efforts. This work was funded by a grant from the National Library of Medicine (R00LM011393).

Ethics Statement

This study was granted an exemption from review by the University of Utah Institutional Review Board (IRB_00076188). Note that in order to protect tweeter anonymity, we have not reproduced tweets verbatim. Example tweets shown were generated by the researchers as exemplars only.

References

American Psychiatric Association. 2000. *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision (DSM-IV-TR)*. American Psychiatric Association.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)*. American Psychiatric Publishing.

J Ayers, B Althouse, and M Dredze. 2014. Could behavioral medicine lead the web data revolution? *JAMA*, 311(14):1399–400, Apr.

L Baer, D G Jacobs, J Meszler-Reizes, M Blais, M Fava, R Kessler, K Magruder, J Murphy, B Kopans, P Cukor, L Leahy, and J O'Laughlen. 2000. Development of a brief screening instrument: the HANDS. *Psychother Psychosom*, 69(1):35–41.

Centers for Disease Control. 2014. *BRFSS - Anxiety and Depression Optional Module*.

M Chary, N Genes, A McKenzie, and A Manini. 2013. Leveraging social networks for toxicovigilance. *J Med Toxicol*, 9(2):184–91, Jun.

C Chew and G Eysenbach. 2010. Pandemics in the age of Twitter: content analysis of tweets during the 2009 H1N1 outbreak. *PLoS One*, 5(11):e14118.

M Conway, A Kawazoe, H Chanlekha, and N Collier. 2010. Developing a disease outbreak event corpus. *J Med Internet Res*, 12(3):e43.

G Coppersmith, M Dredze, and C Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

M De Choudhury, S Counts, and E Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 47–56. ACM.

M Dredze. 2012. How social media will change public health. *Intelligent Systems, IEEE*, 27(4):81–84.

C Freifeld, J Brownstein, C Menone, W Bao, R Filice, T Kass-Hout, and N Dasgupta. 2014. Digital drug safety surveillance: monitoring pharmaceutical products in Twitter. *Drug Saf*, 37(5):343–50, May.

S Gilman, N Trinh, J Smoller, M Fava, J Murphy, and J Breslau. 2013. Psychosocial stressors and the prognosis of major depression: a test of Axis IV. *Psychol Med*, 43(2):303–16, Feb.

C Hanson, B Cannon, S Burton, and C Giraud-Carrier. 2013. An exploration of social circles and prescription drug abuse through Twitter. *J Med Internet Res*, 15(9):e189.

C Homan, R Johar, T Liu, M Lytle, V Silenzio, and C Ovesdotter Alm. 2014. Toward macro-insights for suicide prevention: analyzing fine-grained distress at scale. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 107–117, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

G Hripcsak and A Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *JAMIA*, 12(3):296–298.

S Huang, P LePendu, S Iyer, M Tai-Seale, D Carrell, and N Shah. 2014. Toward personalizing treatment for depression: predicting diagnosis and severity. *J Am Med Inform Assoc*, 21(6):1069–75, Nov.

J Jashinsky, S Burton, C Hanson, J West, C Giraud-Carrier, M Barnes, and T Argyle. 2014. Tracking suicide risk factors through Twitter in the US. *Crisis*, 35(1):51–9.

R Kessler and P Wang. 2009. Handbook of Depression. chapter Epidemiology of Depression, pages 5–22. Guilford Press, 2nd edition.

R Kessler, P Berglund, O Demler, R Jin, D Koretz, K Merikangas, A Rush, E Walters, P Wang, and National Comorbidity Survey Replication. 2003. The epidemiology of major depressive disorder: results from the national comorbidity survey replication (NCS-R). *Journal of the American Medical Association*, 289(23):3095–105.

J-D Kim, T Ohta, Y Tateisi, and J Tsujii. 2003. Genia corpus–semantically annotated corpus for biotextmining. *Bioinformatics*, 19 Suppl 1:i180–2.

K Kroenke, R Spitzer, and J Williams. 2001. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*, 16(9):606–13, Sep.

D Mowery, P Jordan, J Wiebe, H Harkema, J Dowling, and W Chapman. 2013. Semantic annotation of clinical events for generating a problem list. *AMIA Annu Symp Proc*, 2013:1032–41.

M Myslín, S-H Zhu, W Chapman, and M Conway. 2013. Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. *J Med Internet Res*, 15(8):e174.

M Paul, M Dredze, and D Broniatowski. 2014. Twitter improves influenza forecasting. *PLoS Curr*, 6.

J Pestian, H Nasrallah, P Matykiewicz, A Bennett, and A Leenaars. 2010. Suicide note classification using natural language processing: a content analysis. *Biomed Inform Insights*, 2010(3):19–28, Aug.

J Pestian, P Matykiewicz, and M Linn-Gust. 2012. What's in a note: construction of a suicide note corpus. *Biomed Inform Insights*, 5:1–6.

M Poesio and R Vieira. 1998. A corpus-based investigation of definite description use. *Comput. Linguist.*, 24(2):183–216, June.

K Posner, G Brown, B Stanley, D Brent, K Yershova, M Oquendo, G Currier, G Melvin, L Greenhill, S Shen, and J Mann. 2011. The Columbia-Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *Am J Psychiatry*, 168(12):1266–77, Dec.

L Sawyer Radloff. 1977. The CES-D scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1(3):385–401.

A Roberts, R Gaizauskas, M Hepple, N Davis, G Demetriou, Y Guo, J Kola, I Roberts, A Setzer, A Tapuria, and B Wheeldin. 2007. The CLEF corpus: semantic annotation of clinical text. *AMIA Annu Symp Proc*, pages 625–9.

A Rush, M Trivedi, H Ibrahim, T Carmody, B Arnow, D Klein, J Markowitz, P Ninan, S Kornstein, R Manber, M Thase, J Kocsis, and M Keller. 2003. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol Psychiatry*, 54(5):573–83, Sep.

B South, S Shen, J Leng, T Forbush, S DuVall, and W Chapman. 2012. A prototype tool set to support machine-assisted annotation. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, BioNLP '12, pages 130–139, Stroudsburg, PA, USA. Association for Computational Linguistics.

US Burden of Disease Collaborators. 2013. The state of US health, 1990-2010: burden of diseases, injuries, and risk factors. *JAMA*, 310(6):591–608, Aug.

V Vincze, G Szarvas, R Farkas, G Móra, and J Csirik. 2008. The Bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9 Suppl 11:S9.

L Xuan, I Lancashire, G Hirst, and R Jokel. 2011. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. *Literary and Linguistic Computing*, 26(435-461).