Profiling depression and anorexia in Social Media

David E. Losada

CITUS USC

ED profiling
depression
Eating Disorders
ethics
acks
depression

common mental disorder

persistent sadness  self-harm
hopelessness  loss of energy
loss of interest  restlessness  indecisiveness
sleeping problems  change of appetite  suicide
worthlessness
inability to carry out daily activities
reduced concentration  anxiety

World Health Organization

key facts
300 MILLION people
ALL ages
more women than men
800k people die due to suicide

suicide: 2nd leading cause of death in 15-29 year-olds
leading cause of disability

▼ 50% receive treatment
(in many countries ▼ 10%)
and the burden of depression is on the rise
profiling depression

online profiling
as a complementary tool

lack of access to qualified assessments
report inaccurately or underreport symptoms
e.g. to avoid negative consequences
(active duty soldiers, child custody evaluations)
key components

- **text/content analysis**: extraction of symptoms, detection of depression, doc search, ...
- **network/phone usage** statistics
- **resources/data**: LIWC, ontologies, discussion boards, support communities, questionnaires, ...
profiling depression - refs

construction of a depression lexicon (assisted by experts)

to evaluate the level of depression in texts

harvest the web for metaphorical relations in which depression is embedded (e.g. "Depression is like Y")

extracts relevant concepts related to depression
consultation records

semantic-based approach extracts depressive symptoms (depressed mood, suicide ideas, anxiety, ...)

concept hierarchies, Hamilton Depression Rating Scale (HDRS) KBs (HowNet), domain ontology

Experiments done with PsychPark, a virtual psychiatric clinic maintained by a group of volunteer professionals
Hamilton Depression Rating Scale (HAM-D)

(To be administered by a health care professional)

Patient Name ____________________________________________

Today’s Date ________________

The HAM-D is designed to rate the severity of depression in patients. Although it contains 21 areas, calculate the patient’s score on the first 17 answers.

1. DEPRESSED MOOD
   (Gloomy attitude, pessimism about the future, feeling of sadness, tendency to weep)
   0 = Absent
   1 = Sadness, etc.
   2 = Occasional weeping
   3 = Frequent weeping
   4 = Extreme symptoms

6. INSOMNIA - Delayed
   (Waking in early hours of the morning and unable to fall asleep again)
   0 = Absent
   1 = Occasional
   2 = Frequent

7. WORK AND INTERESTS
   0 = No difficulty
   1 = Feelings of incapacity, listlessness, indecision and vacillation
   2 = Loss of interest in hobbies, decreased social activities
   3 = Productivity decreased
   4 = Unable to work. Stopped working because of present illness only. (Absence from work after treatment or recovery may rate a lower score).
consultation records

**semantic-based** approach extracts depressive **symptoms** (depressed mood, suicide ideas, anxiety, ...)

**concept hierarchies**, Hamilton Depression Rating Scale (HDRS) **KBs** (HowNet), domain **ontology**

Experiments done with PsychPark, a **virtual psychiatric clinic** maintained by a group of volunteer professionals
search technology to assist individuals to locate docs related to their depressive problems

consultation docs (long) query (depressive problems & symptoms) recommendations (suggestions & advice written by experts)

high-level discourse analysis
3 main discourse units: events, symptoms & relations

online discussion boards (webmd), SA-UK (www.social-anxiety-community.org/db), John Tung Foundation, www.jtf.org.tw), email databases (www.psychpark.org), HDRS
tweet classification

strongly concerning  possibly concerning  safe to ignore

text classification:  
SVM/Logistic Regression
unigrams + freq-based features

training data:  extracted tweets using pre-defined phrases
and the retrieved tweets were coded by humans

classifier performance:  80%

some individuals broadcast their suicidality on SM
suicide prevention tool?
publicly available profile updates seeking for traces of depression manual coders review histories of status updates according to established clinical criteria Diagnostic and Statistical Manual (DSM) users categorized according to DSM relationship between depression on profile and age, graduation year, gender, relationship status, facebook activity, ...
publicly available profile updates

seeking for traces of depression

manual coders review histories of status updates according to established clinical criteria
Diagnostic and Statistical Manual (DSM)

users categorized according to DSM

relationship between depression on profile and age, graduation year, gender, relationship status, facebook activity, ...
Major Depression Episodes (MDE): symptoms include depressed mood, loss of interest/pleasure in activities, appetite changes, sleep problems, psychomotor agitation/retardation, energy loss, feeling worthless or guilty, decreased concentration or suicidal ideation.

5 or more of these symptoms during the same 2 week period and at least one must be depressed mood or lost of interest/pleasure.
publicly available profile updates seeking for traces of depression

manual coders review histories of status updates according to established clinical criteria Diagnostic and Statistical Manual (DSM)

users categorized according to DSM

relationship between depression on profile and age, graduation year, gender, relationship status, facebook activity, ...
depression    bipolar disorder
seasonal affective disorder (SAD)
post-traumatic stress disorder (PTSD)

automatically identifying self-expressions of mental illness diagnoses
e.g. “I was diagnosed with X.”

statistical classifiers to distinguish each group from a control group
different types of features (LIWC, n-grams, ...)

open-vocabulary analysis: language use relevant to mental health
crowdsourcing

AMT turkers: asked to take a standard clinical depression survey
Beck Depression Inventory (BDI)

followed by self-reported info
and got the turkers' Twitter usernames!

measures of depressive behaviour (engagement, emotion, linguistic style, depression language, activity, ...)

classification powered by different types of features
(emotion from LIWC, time, linguistic style, n-grams, user engagement and ego-network)
Beck's Depression Inventory
This depression inventory can be self-scored. The scoring scale is at the end of the questionnaire.

1. 0 I do not feel sad.
    1 I feel sad
    2 I am sad all the time and I can't snap out of it.
    3 I am so sad and unhappy that I can't stand it.

2. 0 I am not particularly discouraged about the future.
    1 I feel discouraged about the future.
    2 I feel I have nothing to look forward to.
    3 I feel the future is hopeless and that things cannot improve.

3. 0 I do not feel like a failure.
    1 I feel I have failed more than the average person.
    2 As I look back on my life, all I can see is a lot of failures.
    3 I feel I am a complete failure as a person.

4. 0 I get as much satisfaction out of things as I used to.
    1 I don't enjoy things the way I used to.
    2 I don't get real satisfaction out of anything anymore.
    3 I am dissatisfied or bored with everything.

5. 0 I don't feel particularly guilty
    1 I feel guilty a good part of the time.
    2 I feel quite guilty most of the time.
    3 I feel guilty all of the time.

6. 0 I don't feel I am being punished.
    1 I feel I may be punished.
    2 I expect to be punished.
    3 I feel I am being punished.

7. 0 I don't feel disappointed in myself.
crowdsourcing

AMT turkers: asked to take a standard **clinical depression survey** Beck Depression Inventory (BDI)

followed by **self-reported info** and got the turkers' Twitter usernames!

**measures of depressive behaviour** (engagement, emotion, linguistic style, depression language, activity, ...)

**classification** powered by different types of features (emotion from LIWC, time, linguistic style, n-grams, user engagement and ego-network)
crowdsourcing

AMT turkers: asked standard clinical Beck Depression

followed by self-reported info and got the turkers' Twitter usernames!

measures of depressive behaviour (engagement, emotion, linguistic style, depression language, activity, ...)

classification powered by different types of features (emotion from LIWC, time, linguistic style, n-grams, user engagement and ego-network)
crowdsourcing

AMT turkers: asked to take a standard **clinical depression survey**
**Beck Depression Inventory (BDI)**

followed by **self-reported info**
and got the turkers' Twitter usernames!

**measures of depressive behaviour** (engagement, emotion, linguistic style, depression language, activity, ...)

**classification** powered by different types of features
(emotion from LIWC, time, linguistic style, n-grams, user engagement and ego-network)
sample of **tweets** about depression

categorization of tweets:

also performed a **screening test**
(69 young adults):

✓ surveying users (self-judged depression level, CES-D test)
✓ collecting tweets of the same users
✓ comparing depression levels vs sentiments & language
Figure 1: Labeled and categorized language usage in relation to depression
sample of tweets about depression

categorization of tweets:

also performed a screening test (69 young adults):

surveying users (self-judged depression level, CES-D test)

collecting tweets of the same users

comparing depression levels vs sentiments & language
sample of *tweets* about depression

categorization of tweets:

The CESD-R is a screening test for depression and depressive disorder. The CESD-R measures symptoms defined by the American Psychiatric Association's Diagnostic and Statistical Manual (DSM-V) for a major depressive episode. At the top of each of the following screens, you will see a statement. For each statement, please indicate how often you have felt this way recently by selecting the option you most agree with.

*collecting tweets of the same users*
*comparing depression levels vs sentiments & language*
A sample of tweets about depression

categorization of tweets:

![Chart showing categorized language usage in relation to depression]

Depressive Moods of Users Portrayed in Twitter

Minsu Park
KAIST
373-1 Guseong-dong
Deajeon, Korea
mansumansu@kaist.ac.kr

Chiyung Cha
Ewha Womans University
82-2 Daehyeon-dong
Seoul, Korea
chiyoung@ewha.ac.kr

Meeyoung Cha
KAIST
373-1 Guseong-dong
Deajeon, Korea
meeyoungcha@kaist.ac.kr

also performed a screening test
(69 young adults):

/ surveying users (self-judged depression level, CES-D test)

--> collecting tweets of the same users

□ comparing depression levels vs sentiments & language
recruited participants (screening survey)

- demographic data
- Twitter ID
- depression quotient (CES-D)

💡 personal interviews (e.g. at cafe's) participants' experiences with SM experience with depression

📝 coded the (recorded) interviews according to different themes

👥 also got and analysed tweets from the participants' friends
216 college students

CES-D questionnaires

30% met min. criteria for depression

network data (campus) usage statistics contents not recorded!

studied increment of internet usage, avg packets per user, p2p statistics, and compares them among groups
216 college students

CES-D question

30% met min. for depression

network data (corpus)
usage statistic contents not recorded!

studied increment of internet usage, avg packets per user, p2p statistics, and compares them among groups

flows, packets, octets, durations, protocols (chats, p2p, email,...)
user's phone data

Patient Health Questionnaire-9

physical context-motion, variability of location, variability of time, home stay, social settings, phone usage (e.g. screen state)

depressed users visited fewer locations, spent more time at home, moved less through geographic space, had greater phone usage duration and frequency
### PATIENT HEALTH QUESTIONNAIRE-9 (PHQ-9)

Over the last 2 weeks, how often have you been bothered by any of the following problems? (Use "*" to indicate your answer)

<table>
<thead>
<tr>
<th>Problem</th>
<th>Not at all</th>
<th>Several days</th>
<th>More than half the days</th>
<th>Nearly every day</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Little interest or pleasure in doing things</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2. Feeling down, depressed, or hopeless</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3. Trouble falling or staying asleep; or sleeping too much</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4. Feeling tired or having little energy</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>5. Poor appetite or overeating</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>6. Feeling bad about yourself — or that you are a failure or have let yourself or your family down</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>7. Trouble concentrating on things, such as reading the newspaper or watching television</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>8. Moving or speaking so slowly that other people could have noticed? Of the opposite — being so fidgety or restless that you have been moving around a lot more than usual</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>9. Thoughts that you would be better off dead or of hurting yourself in some way</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

For office use only: 0 * 0 * 0 * Total Score: 

If you checked off any problems, how difficult have these problems made it for you to do your work, take care of things at home, or get along with other people?

<table>
<thead>
<tr>
<th>Difficulty Level</th>
<th>Not difficult at all</th>
<th>Somewhat difficult</th>
<th>Very difficult</th>
<th>Extremely difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>
user's phone data

Patient Health Questionnaire-9

physical context-motion, variability of location, variability of time, home stay, social settings, phone usage (e.g. screen state)

depressed users visited fewer locations, spent more time at home, moved less through geographic space, had greater phone usage duration and frequency
essays by college students

- currently-depressed
- formerly-depressed
- never-depressed

Linguistic Inquiry and Word Count (LIWC)

depressed participants:
△ negatively valenced words, neg emotions
△ 1st person singular (I, me, my) (think a great deal about themselves)
△ slightly more positive emotions than never-depressed
### Table 1. LIWC2015 Output Variable Information

<table>
<thead>
<tr>
<th>Category</th>
<th>Abbrev</th>
<th>Examples</th>
<th>Words in category</th>
<th>Internal Consistency (Uncorrected)</th>
<th>Internal Consistency (Corrected)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word count</td>
<td>WC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summary Language Variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analytical thinking</td>
<td>Analytic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clear</td>
<td>Clear</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authentic</td>
<td>Authentic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotional tone</td>
<td>Tone</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words per sentence</td>
<td>WPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words + syllables</td>
<td>Syll</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syntactic words</td>
<td>Synt</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of words</td>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total pronouns</td>
<td>Pronoun</td>
<td>I, me, my, mine</td>
<td>491</td>
<td>.89</td>
<td>.74</td>
</tr>
<tr>
<td>Personal pronouns</td>
<td>Personal</td>
<td>I, me, he, his</td>
<td>135</td>
<td>.75</td>
<td>.67</td>
</tr>
<tr>
<td>3rd person singular</td>
<td>3rd</td>
<td>he, him</td>
<td>55</td>
<td>.85</td>
<td>.81</td>
</tr>
<tr>
<td>3rd person plural</td>
<td>3rd</td>
<td>they, them</td>
<td>24</td>
<td>.76</td>
<td>.71</td>
</tr>
<tr>
<td>Articles</td>
<td>Article</td>
<td>a, an, the</td>
<td>59</td>
<td>.86</td>
<td>.71</td>
</tr>
<tr>
<td>Prepositions</td>
<td>Prepos</td>
<td>in, to, with, above</td>
<td>17</td>
<td>.94</td>
<td>.77</td>
</tr>
<tr>
<td>Auxiliary verbs</td>
<td>AuxVerb</td>
<td>will, would, have</td>
<td>11</td>
<td>.87</td>
<td>.73</td>
</tr>
<tr>
<td>Nouns</td>
<td>Noun</td>
<td>this, those</td>
<td>17</td>
<td>.40</td>
<td>.81</td>
</tr>
<tr>
<td>Adjectives</td>
<td>Adjective</td>
<td>good, beautiful</td>
<td>59</td>
<td>.76</td>
<td>.71</td>
</tr>
<tr>
<td>Adverb</td>
<td>Adverb</td>
<td>so, much</td>
<td>73</td>
<td>.82</td>
<td>.71</td>
</tr>
<tr>
<td>Interrogatives</td>
<td>Interrog</td>
<td>what, why, when</td>
<td>73</td>
<td>.85</td>
<td>.72</td>
</tr>
<tr>
<td>Interjections</td>
<td>Interj</td>
<td>oh, ugh!</td>
<td>73</td>
<td>.85</td>
<td>.72</td>
</tr>
<tr>
<td>Psychological Processes</td>
<td>Psychological</td>
<td>happy, cried</td>
<td>73</td>
<td>.85</td>
<td>.72</td>
</tr>
<tr>
<td>Positive emotion</td>
<td>Positive</td>
<td>love, nice, sweet</td>
<td>620</td>
<td>.72</td>
<td>.64</td>
</tr>
<tr>
<td>Negative emotion</td>
<td>Negative</td>
<td>hate, kill, anger</td>
<td>53</td>
<td>.72</td>
<td>.64</td>
</tr>
<tr>
<td>Anxiety</td>
<td>Anxiety</td>
<td>worried, fearful</td>
<td>156</td>
<td>.87</td>
<td>.71</td>
</tr>
<tr>
<td>Affect</td>
<td>Affect</td>
<td>anger, hate, kill, annoyed</td>
<td>73</td>
<td>.85</td>
<td>.72</td>
</tr>
<tr>
<td>Subsets</td>
<td>Subset</td>
<td>crying, angry, sad</td>
<td>73</td>
<td>.85</td>
<td>.72</td>
</tr>
<tr>
<td>Social processes</td>
<td>Social</td>
<td>own, family</td>
<td>73</td>
<td>.85</td>
<td>.72</td>
</tr>
<tr>
<td>Family</td>
<td>Family</td>
<td>daughter, dad, aunt</td>
<td>73</td>
<td>.85</td>
<td>.72</td>
</tr>
</tbody>
</table>

### LIWC2015 Development Manual

<table>
<thead>
<tr>
<th>Category</th>
<th>Abbrev</th>
<th>Examples</th>
<th>Words in category</th>
<th>Internal Consistency (Uncorrected)</th>
<th>Internal Consistency (Corrected)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friends</td>
<td>friend</td>
<td>buddy, neighbor</td>
<td>85</td>
<td>.79</td>
<td>.64</td>
</tr>
<tr>
<td>Female references</td>
<td>Female</td>
<td>girl, her, mom</td>
<td>124</td>
<td>.51</td>
<td>.87</td>
</tr>
<tr>
<td>Male references</td>
<td>Male</td>
<td>boy, his, dad</td>
<td>116</td>
<td>.53</td>
<td>.87</td>
</tr>
<tr>
<td>Cognitive processes</td>
<td>Cognitive</td>
<td>cause, know, eng</td>
<td>767</td>
<td>.60</td>
<td>.90</td>
</tr>
<tr>
<td>Emotions</td>
<td>Emotions</td>
<td>love, fear</td>
<td>259</td>
<td>.43</td>
<td>.84</td>
</tr>
<tr>
<td>Communication</td>
<td>Communication</td>
<td>because, effect</td>
<td>120</td>
<td>.29</td>
<td>.64</td>
</tr>
<tr>
<td>Disagreement</td>
<td>Disagree</td>
<td>should, wouldn't</td>
<td>15</td>
<td>.94</td>
<td>.74</td>
</tr>
<tr>
<td>Transitive</td>
<td>Transitive</td>
<td>maybe, probably</td>
<td>120</td>
<td>.41</td>
<td>.84</td>
</tr>
<tr>
<td>Certainty</td>
<td>Certainty</td>
<td>certain, always</td>
<td>115</td>
<td>.33</td>
<td>.71</td>
</tr>
<tr>
<td>Discrimination</td>
<td>Discrimination</td>
<td>how's, but also</td>
<td>25</td>
<td>.94</td>
<td>.84</td>
</tr>
<tr>
<td>Hospitability</td>
<td>Hospitable</td>
<td>look, basic, feeling</td>
<td>430</td>
<td>.17</td>
<td>.51</td>
</tr>
<tr>
<td>See</td>
<td>See</td>
<td>view, see, seen</td>
<td>120</td>
<td>.46</td>
<td>.84</td>
</tr>
<tr>
<td>Hear</td>
<td>Hear</td>
<td>listen, hearing</td>
<td>65</td>
<td>.25</td>
<td>.64</td>
</tr>
<tr>
<td>Feel</td>
<td>Feel</td>
<td>feel, touch</td>
<td>120</td>
<td>.24</td>
<td>.64</td>
</tr>
<tr>
<td>Biological processes</td>
<td>Biological</td>
<td>eat, brain, pain</td>
<td>708</td>
<td>.29</td>
<td>.71</td>
</tr>
<tr>
<td>Body</td>
<td>Body</td>
<td>check, hands, arm</td>
<td>215</td>
<td>.52</td>
<td>.8</td>
</tr>
<tr>
<td>Health</td>
<td>Health</td>
<td>health, illness</td>
<td>254</td>
<td>.69</td>
<td>.31</td>
</tr>
<tr>
<td>Perception</td>
<td>Perception</td>
<td>notice, feel, touch</td>
<td>115</td>
<td>.55</td>
<td>.76</td>
</tr>
<tr>
<td>Perception</td>
<td>Perception</td>
<td>notice, feel, touch</td>
<td>115</td>
<td>.55</td>
<td>.76</td>
</tr>
<tr>
<td>Perception</td>
<td>Perception</td>
<td>notice, feel, touch</td>
<td>115</td>
<td>.55</td>
<td>.76</td>
</tr>
<tr>
<td>Virtue</td>
<td>Virtue</td>
<td>direct, indirect</td>
<td>115</td>
<td>.39</td>
<td>.84</td>
</tr>
<tr>
<td>Affiliation</td>
<td>Affiliation</td>
<td>affiliation, self, friend, social</td>
<td>24</td>
<td>.45</td>
<td>.84</td>
</tr>
<tr>
<td>Achievement</td>
<td>Achievement</td>
<td>achieve, win, success, loss</td>
<td>25</td>
<td>.45</td>
<td>.84</td>
</tr>
<tr>
<td>Power</td>
<td>Power</td>
<td>power, superior, bully</td>
<td>51</td>
<td>.51</td>
<td>.74</td>
</tr>
<tr>
<td>Reward</td>
<td>Reward</td>
<td>earn, prize, benefit</td>
<td>25</td>
<td>.26</td>
<td>.64</td>
</tr>
<tr>
<td>Risk</td>
<td>Risk</td>
<td>risk, danger, doubt</td>
<td>163</td>
<td>.26</td>
<td>.64</td>
</tr>
<tr>
<td>Time orientations</td>
<td>TimeOrientation</td>
<td>past tense, future tense, ago, later, before</td>
<td>540</td>
<td>.21</td>
<td>.94</td>
</tr>
<tr>
<td>Present focus</td>
<td>PresentFocus</td>
<td>present tense, today, now</td>
<td>424</td>
<td>.24</td>
<td>.94</td>
</tr>
<tr>
<td>Future focus</td>
<td>FutureFocus</td>
<td>future tense, may, will, seem</td>
<td>65</td>
<td>.26</td>
<td>.64</td>
</tr>
<tr>
<td>Relativization</td>
<td>Relativization</td>
<td>some, based, and</td>
<td>952</td>
<td>.63</td>
<td>.84</td>
</tr>
<tr>
<td>Motion</td>
<td>Motion</td>
<td>move, move</td>
<td>125</td>
<td>.35</td>
<td>.76</td>
</tr>
<tr>
<td>Space</td>
<td>Space</td>
<td>down, on, in, to</td>
<td>460</td>
<td>.45</td>
<td>.81</td>
</tr>
<tr>
<td>Time</td>
<td>Time</td>
<td>time, until, season</td>
<td>310</td>
<td>.39</td>
<td>.71</td>
</tr>
<tr>
<td>Verbal events</td>
<td>VerbalEvent</td>
<td>work, job, move, event</td>
<td>443</td>
<td>.68</td>
<td>.97</td>
</tr>
<tr>
<td>Leisure</td>
<td>Leisure</td>
<td>leisure, cool, chat, movie</td>
<td>296</td>
<td>.55</td>
<td>.83</td>
</tr>
<tr>
<td>Home</td>
<td>Home</td>
<td>home, house, land</td>
<td>140</td>
<td>.46</td>
<td>.83</td>
</tr>
<tr>
<td>Money</td>
<td>Money</td>
<td>money, cash, owe</td>
<td>220</td>
<td>.69</td>
<td>.94</td>
</tr>
<tr>
<td>Religion</td>
<td>Religion</td>
<td>church, faith</td>
<td>174</td>
<td>.64</td>
<td>.94</td>
</tr>
<tr>
<td>Death</td>
<td>Death</td>
<td>death, bury, coffin</td>
<td>54</td>
<td>.55</td>
<td>.75</td>
</tr>
<tr>
<td>Informal language</td>
<td>InformalLanguage</td>
<td>informal, don't, don't</td>
<td>370</td>
<td>.46</td>
<td>.83</td>
</tr>
<tr>
<td>Slang words</td>
<td>SlangWord</td>
<td>bad, down, shit</td>
<td>120</td>
<td>.41</td>
<td>.84</td>
</tr>
<tr>
<td>Slang</td>
<td>Slang</td>
<td>bad, down, shit</td>
<td>120</td>
<td>.41</td>
<td>.84</td>
</tr>
<tr>
<td>Accept</td>
<td>Accept</td>
<td>accept, agree, OK, yes</td>
<td>36</td>
<td>.10</td>
<td>.38</td>
</tr>
<tr>
<td>Nonformal</td>
<td>Nonformal</td>
<td>er, um, umm</td>
<td>19</td>
<td>.37</td>
<td>.69</td>
</tr>
<tr>
<td>Fillers</td>
<td>Fillers</td>
<td>filler, you know</td>
<td>14</td>
<td>.06</td>
<td>.27</td>
</tr>
</tbody>
</table>
essays by college students
- currently-depressed
- formerly-depressed
- never-depressed

Linguistic Inquiry and Word Count (LIWC)

depressed participants:
▲ negatively valenced words, neg emotions
▲ 1st person singular (I, me, my)
  (think a great deal about themselves)
▲ slightly more positive emotions than never-depressed
Using Topic Modeling to Improve Prediction of Neuroticism and Depression in College Students

Philip Resnik
University of Maryland
College Park, MD 20742
resnik@umd.edu

Anderson Garron
University of Maryland
College Park, MD 20742
agarron@cs.umd.edu

Rebecca Resnik
Mindwell Psychology Bethesda
5602 Shields Drive
Bethesda, MD 20817
drrebeccaresnik@gmail.com

esays by college students

BDI score

linear regression

Linguistic Inquiry and Word Count (LIWC) and Latent Dirichlet Allocation (LDA) features

Qualitative analysis: LDA extracts topics whose associated words are themes related to depression
Detecting depression and mental illness on social media: an integrative review
Sharath Chandra Guntuku¹, David B Yaden¹, Margaret L Kern², Lyle H Ungar¹ and Johannes C Eichstaedt¹
¹University of Pennsylvania, Philadelphia, PA, United States
²The University of Melbourne, Melbourne, Australia
Current Opinion in Behavioral Sciences 2017, 18:43–49
Table 1

Prediction performances achieved by different mental illness studies reviewed in this paper. The relevant dataset, features, and prediction settings are provided.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Year</th>
<th>Dataset</th>
<th>Platform</th>
<th>N (Users)</th>
<th>Cases (condition; base rate)</th>
<th>Section</th>
<th>Mental Illness Criteria</th>
<th>Features (predictors)</th>
<th>Outcome Type</th>
<th>Model</th>
<th>Metric</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>[3]</td>
<td>2013</td>
<td>Twitter</td>
<td>476</td>
<td>A</td>
<td>Depression = 171 (BR = 65%)</td>
<td>Survey</td>
<td>(CESD + IDH)</td>
<td>Y Y Y Social Network</td>
<td>Binary</td>
<td>PCA, SVM w/ RBF Kernel</td>
<td>Accuracy</td>
<td>.72</td>
</tr>
<tr>
<td>[14]</td>
<td>2014</td>
<td>Facebook</td>
<td>28,749</td>
<td>A</td>
<td>Continuous Depression score</td>
<td>Survey</td>
<td>Personality</td>
<td>Y Y Y</td>
<td>Continuous Ridge Regression</td>
<td>Correlation</td>
<td>.38</td>
<td></td>
</tr>
<tr>
<td>[12]</td>
<td>2015</td>
<td>Twitter</td>
<td>209</td>
<td>A</td>
<td>Depression = 83 (BR = 35%)</td>
<td>Survey</td>
<td>(CBSU)</td>
<td>Y Y Y User Activity</td>
<td>Binary</td>
<td>SVM</td>
<td>Accuracy</td>
<td>.69</td>
</tr>
<tr>
<td>[40]</td>
<td>2014</td>
<td>Twitter</td>
<td>5,972</td>
<td>B</td>
<td>FTSD = 244 (BR = 4%)</td>
<td>Self-declared</td>
<td>Y Y Y</td>
<td>Binary (not reported)</td>
<td>ROC</td>
<td>(AUC not reported)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[42]</td>
<td>2014</td>
<td>Twitter</td>
<td>21,866</td>
<td>B</td>
<td>11,866 (across 4 Conditions, BR = 54%)</td>
<td>Self-declared</td>
<td>Y Y Y Y User Activity</td>
<td>Binary Logistic Regression</td>
<td>Precision*</td>
<td>Depression = .48</td>
<td>Bipolar = .64</td>
<td>FTSD = .67</td>
</tr>
<tr>
<td>[17]</td>
<td>2015</td>
<td>Twitter</td>
<td>1,937</td>
<td>B</td>
<td>Depression = 481 (BR = 15%)</td>
<td>Self-declared</td>
<td>Y Y Y Age, Gender, Personality</td>
<td>Binary Logistic Regression</td>
<td>AUC</td>
<td>Depression = .81</td>
<td>FTSD = .48</td>
<td></td>
</tr>
<tr>
<td>[21]</td>
<td>2015</td>
<td>Twitter</td>
<td>4,026</td>
<td>B</td>
<td>7,013 (across 10 Conditions, BR = 50%)</td>
<td>Self-declared</td>
<td>Y Y Y</td>
<td>Binary (not reported)</td>
<td>Precision*</td>
<td>Depression = .48</td>
<td>Bipolar = .43</td>
<td>Anxiety = .45</td>
</tr>
<tr>
<td>[14]</td>
<td>2016</td>
<td>Twitter</td>
<td>250</td>
<td>B</td>
<td>Suicide Attempt = 125 (BR = 50%)</td>
<td>Self-declared</td>
<td>Y Y Y User Activity</td>
<td>Binary (not reported)</td>
<td>Precision*</td>
<td></td>
<td>.70</td>
<td></td>
</tr>
<tr>
<td>[43]</td>
<td>2016</td>
<td>Twitter</td>
<td>900</td>
<td>B</td>
<td>Depression = 326 (BR = 16%)</td>
<td>Self-declared</td>
<td>Y</td>
<td>Binary Noi-Bayer</td>
<td>AUC</td>
<td></td>
<td>.70</td>
<td></td>
</tr>
<tr>
<td>[19]</td>
<td>2017</td>
<td>Twitter</td>
<td>9,611</td>
<td>B</td>
<td>MDD (across 8 Conditions, BR = 54%)</td>
<td>Self-declared</td>
<td>Y</td>
<td>Gender Multi-Task Neural Network</td>
<td>AUC</td>
<td>Depression = .76</td>
<td>Bipolar = .55</td>
<td>Depression = .76</td>
</tr>
</tbody>
</table>

AUC: Area Under the Receiver Operating Characteristic (ROC) Curve; Precision: fraction of cases ruled positive that are truly positive; Accuracy: fraction of cases that are correctly labeled by the model; SVM: Support Vector Machines; PCA: Principal Component Analysis; RBF: Radial Basis Function.

*Precision with 10% False Alarms.

aWithin-sample (not cross-validated).

bUsing the Depression facet of the Neuroticism factor measured by the International Personality Item Pool (IPIP) proxy to the NEO-PI-R Personality Inventory [38].

Studies highlighted in green report AUCs: AUCs are not base rate dependent and can be compared across studies.
Detecting depression and mental illness on social media: an integrative review
Sharath Chandra Guntuku¹, David B Yaden¹, Margaret L Kern², Lyle H Ungar¹ and Johannes C Eichstaedt¹

¹ University of Pennsylvania, Philadelphia, PA, United States
² The University of Melbourne, Melbourne, Australia

Current Opinion in Behavioral Sciences 2017, 18:43–49

### Table 1

<table>
<thead>
<tr>
<th>Feature</th>
<th>Number of Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-reported survey</td>
<td>5 studies</td>
</tr>
<tr>
<td>Self-structured on mobile</td>
<td>7 studies</td>
</tr>
<tr>
<td>Person membership</td>
<td>2 studies</td>
</tr>
</tbody>
</table>

### Note

Note: The table above shows the number of studies focused on different aspects of social media use and mental health. The data is aggregated from various sources, including academic journals and reports. Further research is needed to validate these findings.

---

**Figure 1**

Example page from the paper: Detecting depression and mental illness on social media: an integrative review. The figure illustrates the methods used in the study, including data analysis and interpretation. The paper was published in *Current Opinion in Behavioral Sciences* in 2017, volume 18, pages 43–49.
Data sources used in studies as assessment criteria to establish mental illness status. The number of studies selected for review in the present article is provided. The most commonly used self-reported screening surveys for depression include the PHQ-9 = Patient Health Questionnaire [7], CES-D = Centers for Epidemiological Studies Depression Scale Revised [9], BDI = Beck Depression Inventory [10].
Detecting depression and mental illness on social media: an integrative review
Sharath Chandra Guntuku\textsuperscript{1}, David B Yaden\textsuperscript{1}, Margaret L Kern\textsuperscript{2}, Lyle H Ungar\textsuperscript{1} and Johannes C Eichstaedt\textsuperscript{1}

\textsuperscript{1} University of Pennsylvania, Philadelphia, PA, United States
\textsuperscript{2} The University of Melbourne, Melbourne, Australia

Current Opinion in Behavioral Sciences 2017, 18:43–49
Example post: How did this happen to me?

### N-grams

1-grams: How, did, this, happen, to, me, ?  
2-grams: "How did", "did this", "this happen": "happen ?":  
3-grams: "How did this", "did this happen", "this happen ?"  
...

### LIWC (Dictionaries)

<table>
<thead>
<tr>
<th>Category</th>
<th>% Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-references (I, me, my)</td>
<td>16.67</td>
</tr>
<tr>
<td>Social words</td>
<td>0</td>
</tr>
<tr>
<td>Emotions</td>
<td>0</td>
</tr>
<tr>
<td>Overall cognitive words</td>
<td>16.67</td>
</tr>
</tbody>
</table>

### Sentiment

- Subjectivity: polar (0.6)  
- Polarity: negative (0.7)  
- LabMT (Happiness score): -0.21

### N-grams meta-data

Avg. 1-gram length,  
Avg. number of 1-grams per post,  
Total number of 1-grams per user,  
...

### User activity

- Posts by the hour of the day  
- Posts between 12am and 6am  
- Retweets  
- Posts with URLS, Hashtags, @-mentions  
...

### User social network

- Friends  
- Followers  
- People in extended circles  
...

Examples of features included in the different feature sets referenced in Table 1. LIWC: Linguistic Inquiry and Word Count [20], LabMT: Language Assessment by Mechanical Turk [39].

www.sciencedirect.com
Detecting depression and mental illness on social media: an integrative review
Sharath Chandra Guntuku¹, David B Yaden¹, Margaret L Kern², Lyle H Ungar¹ and Johannes C Eichstaedt¹

¹ University of Pennsylvania, Philadelphia, PA, United States
² The University of Melbourne, Melbourne, Australia

Current Opinion in Behavioral Sciences 2017, 18:43–49
Profiling depression and anorexia in Social Media

David E. Losada
eating disorders (ED) are complex mental disorders responsible for the highest mortality rate among mental illnesses.

“Pro-ana” refers to individuals with an ED disorder who focus on having an ED as a lifestyle choice as opposed to a psychiatric disorder.

"Thinspiration" desire to be thin.
sufferers restrict their eating to keep low weight

20% of all deaths from anorexia are the result of suicide

low perceptions of body image

unrealistic ideals of thinness (e.g. based on Internet models)
Bulimia nervosa

repeated cycles of binge eating and purging
eating disorders

symptoms

extreme behavioural/emotional responses to eating food & gaining weight

anxiety

depression

laxative abuse

self-starvation
prevalence of ED has significantly grown

The costs of eating disorders
Social, health and economic impacts
Assessing the impact of eating disorders across the UK on behalf of BEAT.

February 2015
prevalence of ED has significantly grown

The costs of eating disorders
Social, health and economic impacts

Assessing the impact of eating disorders across the UK on behalf of BEAT
February 2011

beat
prevalence of ED has significantly grown

The costs of eating disorders
Social, health and economic impacts
There are increasing numbers of reported cases of eating disorders in the UK

Separately from prevalence data research involving GP data in the UK indicates an increase in the age-standardised annual incidence of all diagnosed eating disorders (for ages 10-49) from 32.3 to 37.2 per 100,000 between 2000 and 2009. This increase appears to be due to an increase in the unspecified eating disorder category as AN and BN numbers remained fairly stable.

Separately, as outlined in Table 3.3, time series analysis of data on the total number of cases of eating disorders being diagnosed in England illustrates a similar trend in increasing prevalence over time with a 34% increase in admissions since 2005-06 – approximately 7% per annum.

These recorded changes may reflect increases in the understanding of eating disorders especially the lesser known disorders and particularly binge eating disorder which has only recently been acknowledged in statistical recording.

<table>
<thead>
<tr>
<th>Year</th>
<th>Count of Finished Admissions Episodes (FAEs) where the primary diagnosis was of eating disorders (England)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005-06</td>
<td>1,882</td>
</tr>
<tr>
<td>2006-07</td>
<td>1,924</td>
</tr>
<tr>
<td>2007-08</td>
<td>1,872</td>
</tr>
<tr>
<td>2008-09</td>
<td>1,886</td>
</tr>
<tr>
<td>2009-10</td>
<td>2,057</td>
</tr>
<tr>
<td>2010-11</td>
<td>[missing data]</td>
</tr>
<tr>
<td>2011-12</td>
<td>2,285</td>
</tr>
<tr>
<td>2012-13</td>
<td>2,380</td>
</tr>
<tr>
<td>2013-14</td>
<td>2,855</td>
</tr>
</tbody>
</table>
prevalence of ED has significantly grown

The costs of eating disorders
Social, health and economic impacts

Assessing the impact of eating disorders across the UK on behalf of BEAT.
February 2011
prevalence of ED has significantly grown

The costs of eating disorders
Social, health and economic impacts
prevalence of ED has significantly grown

The costs of eating disorders
Social, health and economic impacts

Figure 3.3
Gender breakdown of survey respondents

Assessing the impact of eating disorders across the UK on behalf of BEAT
February 2013
prevalence of ED has significantly grown

The costs of eating disorders
Social, health and economic impacts

Assessing the impact of eating disorders across the UK on behalf of BEAT
February 2011
prevalence of ED has significantly grown
prevalence of ED has significantly grown

The costs of eating disorders
Social, health and economic impacts

Assessing the impact of eating disorders across the UK on behalf of BEAT
February 2011
prevalence of ED has significantly grown
clinical studies

surveys & interviews

limitations:

- **small** groups of individuals
- **denial** of illness
- participants **conceal** their condition or its extent
- **ambivalence** towards treatment
- high **drop-out** rates
- predefined **questionnaires** alone may be **insufficient** to reveal the physical/psychological states
Profiling depression and anorexia in Social Media

David E. Losada

ciTUS USC
ED profiling

natural language can be indicative of personality, social status, emotions, mental health, disorders, ...

and interactions/communities in SM can also provide useful signals ...
people's **behaviour** + **content** generated on SM → infer their mental health states

**(semi-)anonymous & open** nature of SM: encourages people to **socialize** and **self-disclose**

naturally occurring data in a non-reactive way.

**SM data** complements **conventional data**
harmful content vs useful recommendations & advice

pro-ED communities (support engagement with ED lifestyles)
people affected by anorexia: age group in which SM are used **heavily**

**contagion** effects on those exposed to dangerous content

pro-ED websites **negatively affect** females’ **perceptions of their body image**

**pro-ED communities:**
- often “hidden in plain sight”
- use of **atypical language or tags**
  (seek to avoid outsiders encountering and reporting them)
pro-anorexia communities:

claim to provide support

promote disordered eating

discourage people from seeking help or trying to recover

- smoke as much as necessary
- use drugs to lose weight
- use laxatives
- eat sugar-free chewing gum
- drink a lot of water to avoid eating
ED & SM: relevant refs

3 doctors

informative → pro-anorexia → others

pro-anorexia info found in **29.3%** of anorexia-related videos

pro-anorexia content: more **highly favored & rated**

**82.6%** of pro-anorexia video raters ☑ liked the misleading info

need to raise awareness about the **trustworthiness of online information**

**Health authorities**: study the content dissemination strategies used by the pro-anorexics & design their own **dissemination strategies for informative content**

**Robust search engines**: find trustworthy content & filter out misleading information
made personal contacts with several pro-ED groups on Facebook and MySpace to get access to, observe and analyze the groups' content

large presence of pro-ana/pro-bulimia groups

低廉 harmful for the viewers/participants....

but also found positive social interactions (social support, help with isolation, ...)

Linguistic Inquiry and Word Count (LIWC) to compare the psychological processes and personal concerns of pro-ED users amongst the two SM sites

Content analysis revealed some differences between the two social networking sites
ED & SM: relevant refs

personal weblogs, a popular form of text-based, diary-like, online journals.

31 pro-ED blogs, 29 recovery blogs, and 27 control blogs

language of pro-ED blogs: lower cognitive processing, a more closed-minded writing style, less emotionally expressive, contained fewer social references, and focused more on eating-related contents than recovery blogs.

12 language indicators correctly classified the blogs in 84% of the cases.

language patterns reflect the psychological conditions of the blog authors and provide insight into their various stages of coping
ED & SM: relevant refs

**pro-ED Twitter profiles'** refs to EDs

45 Pro-ED profiles

how the **followers** reference EDs

profile info + all tweets +
random sample of followers

eexpressions of **disordered** eating patterns & **notable audience of followers**

might provide **social support** but **reinforce an ED identity**
ED & SM: relevant refs

⚠️ Instagram banned searches on several proED tags and issued content advisories on others

investigated pro-ED communities in the aftermath of moderation

despite moderation strategies, pro-ED communities are active and thriving

pro-ED community adopted nonstandard lexical variations of moderated tags to circumvent restrictions

increasingly complex lexical variants emerged over time

more toxic, self-harm, and vulnerable content
ED & SM: relevant refs

analyzed photo sharing on Flickr

is posting of ED content discouraged by posting of recovery-oriented content?

pro-anorexia and pro-recovery communities interact to a high degree

pro-recovery community takes steps to ensure that their content is visible to the pro-anorexia community:

pro-recovery users: employ similar words to those used by pro-anorexia users to describe their photographs comment to pro-anorexia content (counterproductive, entrenches pro-anorexia users in their stance)

242,710 pics

491 users
ED & SM: relevant refs

explored **community structures** and **interactions**
among individuals who suffer from ED

**snowball sampling:** individuals
who self-identify as ED (profile)
+ their connections (followers/followees)

**predictive models:** ED vs non-EDs
(SVM, **97% accuracy**) 📊

Analyzed **social status**, **behavioural patterns** and **psychometric properties**

🔍 **key characteristics of ED:** young ages, prevailing urges to **lose weight** even if being clinically underweight, high social **anxiety**, intensive **self-focused** attention, deep **negative emotion**, increased mental **instability**, and excessive concerns of **body image** and **ingestion** patterns of **homophily** (tendency of individuals to connect with others who share similar characteristics)
ED & SM: relevant refs

*Instagram posted content on pro-ED tags*

LDA topic modelling + novice/clinical annotations

↓

mental illness severity (MIS) in user's content

MIS rating prediction with regression models

forecast MIS levels up to 8 months in the future

% of users whose content expresses high MIS

13%/year increase

2012 2016

26M posts

100k users
ED & SM: relevant refs

**tumblr:** pro-anorexia and pro-recovery communities

**pro-anorexia:**
- enacting anorexia as a **lifestyle** choice
- common **pro-anorexia tags**

**pro-recovery:**
- try to **educate** pro-anorexia individuals of the **health risks** of anorexia

**distinctive** affective, social, cognitive and linguistic style **markers**
- pro-anorexics: greater negative affect, higher cognitive impairment, greater feelings of social isolation and self-harm

**predictive technology:** detect anorexia content (80% accuracy)
content removed
(against community guidelines)
30K pro-ED posts that were public
on Instagram but have since been removed

distinctive signals in deleted content:
more dangerous actions, self-harm tendencies,
and vulnerability (wrt posts that remain public)

classifier: public pro-ED posts vs removed posts (69% acc)

possible applications:
- identify moments for just-in-time intervention
  (e.g. contact a friend or reach out to a specialists)
- facilitate better content moderation
dataset of 1M **photo posts** associated with EDs

**Tumblr**: prohibits the glorification of self-harm, and promoting EDs their accompanying lifestyles

**multimodal**: textual + visual features of pro-ED content

**deep learning** classifier to detect content that violates community guidelines (state-of-the-art Deep Neural Network) **performed comparably to ground truth** that included actually moderated Tumblr data

**Possible application**: pruning the search space of posts that need intervention.
ED & SM: main research themes

impact/prevalence of ED-related contents in SM
analysis of communities & interactions
content analysis, psychometrics
content moderation/violation of SM guidelines
misleading health-related info
effect of moderation
learning technology (classification/regression)
ethics

honor the privacy of the affected individuals

abide by appropriate ethical guidelines

data

use cases

freedom of speech vs security
data

**sensitive/private data**

/ informed consent

ensure security/privacy (storage, access, firewalls)

🚫 no disclosure of personally identifiable info

vs

**public data**

🚫 no interaction with subjects

no need of institutional review approval

avoids the need of contacting subjects, which can be coercive and may change user behaviour
use cases

automatic assessments?
use cases

automatic assessments?
use cases
automatic assessments?

NO WAY!!!
use cases

automatic assessments?

NO WAY!!!

automatic methods CANNOT be used as standalone techniques for diagnosis

need to involve clinicals/psychiatrists
detect possible signs of risk vs clinical assessment
interventions?

- ensure that the intended **benefits** of these interventions outweigh the **risks**

⚠️ **reveal** SM detected **risk to**... the individual himself or an identified trusted social contact or clinician

some SM sites have **basic intervention systems**
interventions?
use cases: example

suggests possible uses of this technology:

// users must **consent to:**
their tweets being **monitored** by an organisation or an individual
**permission to be contacted** if ‘strongly concerning’ tweet detected
use cases: example

suggests possible uses of this technology:

mental health agencies

tool for automatic screening for depression

given the subject’s permission, the system may proactively and automatically screen for signs of depression (e.g. within the subject's online posts)
use cases: example

suggests possible uses of this technology:

    mental health tool for automation for depression

given the subject and automated (e.g. within the

    signs of depression?

    the questionnaire also identifies symptoms of depression?

    the subject is advised to consult a mental health expert

    the subject is informed and offered the opportunity to complete a short online questionnaire
freedom of speech vs security

what contents should be banned by Social Media?

is social media content a lethal threat to vulnerable people?

how effective would the interventions be, in terms of suppressing risk content?
freedom of speech vs security

Advantages of "moderation" policies

- moderating deviant content can constrain sentiments that might harm individuals/communities
- avoids contagion-like effects
- favours user engagement (negative content causes people to leave)

Advantages of "no moderation" policies

- it is better for vulnerable people to identify and express themselves
- discussing dangerous ideas might help people disinhibit themselves from self-harm
- after banning certain contents, some communities became more insular and focused on more dangerous ideas
Advantages of "moderation" policies

- **moderation** deviant content can constrain sentiments that might harm individuals/communities
- avoids contagion-like effects
- favours user engagement (negative content causes people to leave)
freedom of speech vs security

Advantages of "moderation" policies

- moderating deviant content can constrain sentiments that might harm individuals/communities
- avoids contagion-like effects
- favours user engagement (negative content causes people to leave)

Advantages of "no moderation" policies

- it is better for vulnerable people to identify and express themselves
- discussing dangerous ideas might help people disinhibit themselves from self-harm
- after banning certain contents, some communities became more insular and focused on more dangerous ideas
Advantages of "no moderation" policies

- it is better for vulnerable people to identify and express themselves
- discussing dangerous ideas might help people disinhibit themselves from self-harm
- after banning certain contents, some communities became more insular and focused on more dangerous ideas
freedom of speech vs security

Advantages of "moderation" policies

- moderating deviant content can constrain sentiments that might harm individuals/communities
- avoids contagion-like effects
- favours user engagement (negative content causes people to leave)

Advantages of "no moderation" policies

- it is better for vulnerable people to identify and express themselves
- discussing dangerous ideas might help people disinhibit themselves from self-harm
- after banning certain contents, some communities became more insular and focused on more dangerous ideas
freedom of speech vs security

need of collaborations from **industry professionals, researchers, designers, psychologists**, and other **stakeholders** to make decisions in this area....

- avoids contagion-like effects
- favours user engagement (negative content causes people to leave)

- discussing dangerous ideas might help people disinhibit themselves from self-harm
- after banning certain contents, some communities became more **insular** and focused on more **dangerous** ideas
Profiling depression and anorexia in Social Media
David E. Losada
CITUS USC
eRisk

early Risk prediction on the Internet

explores the evaluation methodology, effectiveness metrics and practical applications (particularly those related to health and safety) of early risk detection on the Internet

http://early.irlab.org/
eRisk @ clef 2017

1. **Workshop** open to the submission of papers describing **test collections** or **data sets** suitable for early risk prediction, early risk prediction challenges, tasks and evaluation **metrics** or specific early risk detection **solutions**

2. **Pilot task on early risk detection of depression**: exploratory task on early risk detection of depression, sequentially processing pieces of evidence and detect early traces of depression as soon as possible
1. **Task1. Early Detection of Signs of Depression**  
   (continuation of the eRisk 2017 pilot task)  
   sequentially processing pieces of evidence and detect early traces of depression as soon as possible.

2. **Task2. Early Detection of Signs of Anorexia**  
   (new in 2018)  
   sequentially processing pieces of evidence and detect early traces of anorexia as soon as possible.
data

early intervention is crucial
current technology

doesn´t support early alerts

reactive

works with very explicit signals
current technology

too often, too late!

doesn´t support early alerts

reactive

works with very explicit signals
Data

Key aims

Instigate research on the onset of depression

Proactive technologies

Early alerts

Track temporal evolution
Lack of data on depression & language

few collections available

focus on 2-class categorisation

no temporal dimension, no early risk analysis
little context about the tweet writer
difficult to assess whether a mention of depression is genuine
no way to extract a long history of tweets (e.g. several years)
A Thin Line

no way to extract any history

short messages, little context
large history for each redditor (several years)

many subreddits (communities) about different medical conditions (e.g. depression or anorexia)

long messages

terms & conditions allow use for research purposes
depression group vs control group

Adopted extraction method from Coppersmith et al. 2014:

pattern matching search

search for explicit mentions of diagnosis (e.g. “I was diagnosed with depression”) “I am depressed” “I think I have depression”

manual inspection of the results
depression group vs control group

large set of random redditors

from a wide range of subreddits
(news, media, ...)

also included some false positives
from the depression subreddit
(e.g. “My wife has depression”, “I am a student interested in depression”)
repositor profile

retrieved all history
from any subreddit
his/her posts +
his/her comments to other posts

often several years of text
removed the post/comment with
the explicit mention of the
diagnosis (depression group)
pre- & post-diagnosis text
organised the writings in
chronological order
XML archives
detect early traces of depression

for each subject, sequentially process pieces of evidence...

John Doe’s writings (post or comments)

2/13/13 2/15/13

chunk 1 (oldest writings) 10% of writings

possible case of depression
no depression
no decision yet
detect **early traces** of depression

for each subject, *sequentially process* pieces of evidence...

John Doe’s writings (post or comments)

2/13/13 2/15/13

chunk 1 (oldest writings)
10% of writings

possible case of depression
no depression
no decision yet
detect **early traces** of depression

for each subject, **sequentially process** pieces of evidence...

John Doe’s writings
(post or comments)

2/13/13 2/15/13

chunk 1 (oldest writings)
10% of writings

possible case of depression
no depression
no decision yet
detect **early traces** of depression

for each subject, **sequentially process** pieces of evidence...

John Doe’s writings (post or comments)

<table>
<thead>
<tr>
<th>Date</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/13/13</td>
<td>chunk 1</td>
</tr>
<tr>
<td>2/15/13</td>
<td>chunk 1</td>
</tr>
<tr>
<td>3/3/13</td>
<td>chunk 2</td>
</tr>
<tr>
<td>3/15/13</td>
<td>chunk 2</td>
</tr>
</tbody>
</table>

chunk 1 (oldest writings)
10% of writings

chunk 2
10% of writings

- possible case of depression
- no depression
- no decision yet
detect **early traces** of depression

for each subject, **sequentially process** pieces of evidence...

**John Doe's writings**  
(post or comments)

<table>
<thead>
<tr>
<th>2/13/13</th>
<th>2/15/13</th>
<th>...</th>
<th>1/12/15</th>
<th>3/12/15</th>
</tr>
</thead>
<tbody>
<tr>
<td>chunk 1 (oldest writings) ... 10% of writings</td>
<td>chunk 10 (newest writings) 10% of writings</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

possible case of depression
no depression
no decision yet
Early Risk Detection Error:

\[ \text{ERDE}_o(d,k) = \begin{cases} 
  c_{fp} & \text{(false positive)} \\
  c_{fn} & \text{(false negative)} \\
  c_{tp} \cdot I_c_o(k) & \text{(true positive)} \\
  0 & \text{(true negative)} 
\end{cases} \]
performance metric

True Positive cost

Latency cost function

penalty to late detections
<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Depressed</td>
<td>Control</td>
</tr>
<tr>
<td>Num. subjects</td>
<td>83</td>
<td>403</td>
</tr>
<tr>
<td>Num. submissions (posts &amp; comments)</td>
<td>30,851</td>
<td>264,172</td>
</tr>
<tr>
<td>Avg num. of submissions per subject</td>
<td>371.7</td>
<td>655.5</td>
</tr>
<tr>
<td>Avg num. of days from first to last submission</td>
<td>572.7</td>
<td>626.6</td>
</tr>
<tr>
<td>Avg num. words per submission</td>
<td>27.6</td>
<td>21.3</td>
</tr>
</tbody>
</table>

*Table 1. Main statistics of the train and test collections*
Training

Nov 30th, 2016

all history of all training users provided to the participants (all chunks)
Feb 6th, 2017
chunk1 of all test users provided

Feb 13th, 2017
decisions on chunk1
chunk2 of all test users provided

Feb 20th, 2017
decisions on chunks 1-2
chunk3 of all test users provided

Apr 10th, 2017
decisions on chunks 1-10
<table>
<thead>
<tr>
<th>Institution</th>
<th>Submitted files</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSEEIHT, France</td>
<td>GPLA, GPLB, GPLC, GPLD</td>
</tr>
<tr>
<td>FH Dortmund, Germany</td>
<td>FHDOA, FHDOB, FHDOC, FHDOD, FHDOE</td>
</tr>
<tr>
<td>U. Arizona, USA</td>
<td>UArizonaA, UArizonaB, UArizonaC, UArizonaD, UArizonaE</td>
</tr>
<tr>
<td>U. Nacional de San Luis, Argentina</td>
<td>UNSLA</td>
</tr>
<tr>
<td>U. of Quebec in Montreal, Canada</td>
<td>UQAMA, UQAMB, UQAMC, UQAMD, UQAME</td>
</tr>
<tr>
<td>UACH-INAOE, Mexico-USA</td>
<td>CHEPEA, CHEPEB, CHEPEC, CHEPED</td>
</tr>
<tr>
<td>ISA FRCCSC RAS, Russia</td>
<td>NLPISA</td>
</tr>
<tr>
<td></td>
<td>$ERDE_5$</td>
</tr>
<tr>
<td>--------</td>
<td>----------</td>
</tr>
<tr>
<td>GPLA</td>
<td>17.33%</td>
</tr>
<tr>
<td>GPLB</td>
<td>19.14%</td>
</tr>
<tr>
<td>GPLC</td>
<td>14.06%</td>
</tr>
<tr>
<td>GPLD</td>
<td>14.52%</td>
</tr>
<tr>
<td>FHDOA</td>
<td>12.82%</td>
</tr>
<tr>
<td>FHDDB</td>
<td>12.70%</td>
</tr>
<tr>
<td>FHDCC</td>
<td>13.24%</td>
</tr>
<tr>
<td>FHDOD</td>
<td>13.04%</td>
</tr>
<tr>
<td>FHDOE</td>
<td>14.16%</td>
</tr>
<tr>
<td>UArizonA A</td>
<td>14.62%</td>
</tr>
<tr>
<td>UArizonB</td>
<td>13.07%</td>
</tr>
<tr>
<td>UArizonC</td>
<td>17.93%</td>
</tr>
<tr>
<td>UArizonD</td>
<td>14.73%</td>
</tr>
<tr>
<td>UArizonE</td>
<td>14.93%</td>
</tr>
<tr>
<td>LyRA</td>
<td>15.65%</td>
</tr>
<tr>
<td>LyRB</td>
<td>16.75%</td>
</tr>
<tr>
<td>LyRC</td>
<td>16.14%</td>
</tr>
<tr>
<td>LyRD</td>
<td>14.97%</td>
</tr>
<tr>
<td>LyRE</td>
<td>13.74%</td>
</tr>
<tr>
<td>UNSLA</td>
<td>13.66%</td>
</tr>
<tr>
<td>UQAMA</td>
<td>14.03%</td>
</tr>
<tr>
<td>UQAMB</td>
<td>13.78%</td>
</tr>
<tr>
<td>UQAMC</td>
<td>13.58%</td>
</tr>
<tr>
<td>UQAMD</td>
<td>13.23%</td>
</tr>
<tr>
<td>UQAME</td>
<td>13.68%</td>
</tr>
<tr>
<td>CHEPEA</td>
<td>14.75%</td>
</tr>
<tr>
<td>CHEPEB</td>
<td>14.78%</td>
</tr>
<tr>
<td>CHEPEC</td>
<td>14.81%</td>
</tr>
<tr>
<td>CHEPED</td>
<td>14.81%</td>
</tr>
<tr>
<td>NLPISA</td>
<td>15.59%</td>
</tr>
</tbody>
</table>
THANKS!
funding

This work was supported by the “Ministerio de Economía y Competitividad” of the Government of Spain and FEDER Funds under research projects TIN2012-33867 and TIN2015-64282-R.

This research was funded by the Swiss National Science Foundation (project “Early risk prediction on the Internet: an evaluation corpus”, 2015)
resources

CC BY 2.0: https://creativecommons.org/licenses/by/2.0/

Mark Ingle. (markingleukc at flickr). Trapped. https://goo.gl/PX8hMv
brett jordan. Tell me about your mother board. https://goo.gl/eTKjt5
Matti Mattila. Dotted world map. https://goo.gl/hcfdrN
Simon Cockell. Random, scale-free network. https://goo.gl/BtfW3t
ankxt. Are you ok? https://goo.gl/gKQRu3
Gerald Gabernig. winter.depression. https://goo.gl/xb8ooK
Tim Morgan. database. https://goo.gl/Cy1Ncu
Oscar Rethwill. AH&DY 100m. https://goo.gl/9NK8kF
woodleywonderworks. Pablo's cubism period began at three. https://goo.gl/zhKHF4
resources

CC BY-NC-ND 2.0. https://creativecommons.org/licenses/by-nc-nd/2.0/

collective nouns. https://goo.gl/jWjLCr
resources

CC BY-ND 2.0. https://creativecommons.org/licenses/by-nd/2.0/

Susana Fernandez. Friendship. https://goo.gl/N3Co1W

CC BY-SA 2.0. https://creativecommons.org/licenses/by-sa/2.0/

Helen Harrop. trapped in the shadows. https://goo.gl/odnQp2
resources

CC BY-NC 2.0. https://creativecommons.org/licenses/by-nc/2.0/

WRme2. Grangemouth. https://goo.gl/gibn4w
Timothy Takemoto. The Mirror of the Japanese is not the Gaze of the others. https://goo.gl/1xHdGA
Frits Ahlefeldt Founder Hiking.org. you got hikers illustration. https://goo.gl/cRgPNc
Mark Smiciklas. Social Media ROI. https://goo.gl/qUt4NR
Kathleen Donovan. Chat Bubble. https://goo.gl/mAsBjQ
Andy Kennelly. grouping. https://goo.gl/t8Y1Mh
Emily. The Right Tool. https://goo.gl/EYxtYr