Depression in Social Media: the Experience of the eRisk Lab of CLEF

Prof. Fabio Crestani
Universitá della Svizzera Italiana (USI)
Switzerland
Outline

• The eRisk initiative
  • How we built some test collections
  • Worked on depression, anorexia, and self harm
  • Moving from Early Detection to Level Estimation

• Focus on depression
  • A short summary of what has been done in eRisk so far on this topic
  • An outlook of what is to come
350 million people suffer from Mental Health Problems and, in particular from depression.
early detection and intervention is fundamental
The approach requires: 

**human expert** + **technology**
Current technology does not support early detection. Reactive works only with very explicit signals too often, it works too late!
eRisk is a Lab of CLEF

Investigating the onset of Mental Disorders

with proactive technologies for:

1 - early alerts

2 - tracking temporal evolution

eRisk started in 2017, more info at: https://early.irlab.org/
Text analytics

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

... and we produce a lot of natural language in social media!
The approach is complex

content vs style

linguistic markers

topic models

psychometrics

statistical properties of text

psychological processes

social words

cognitive processes

verb tense

positive/negative emotions

use of personal pronouns
Creation of Test Collections

TREC is a conference and a competition.

Since 2017 we have created 6 test collections for as many different tasks.
Focus on Anxiety and Depression

According to the Anxiety and Depression Association of America (ADAA):

anxiety and depression are the most common mental health problems!

Effective treatment strategies exist and typically involve a combination of talk therapy, medication, and certain lifestyle changes.
Lack of data on depression & language use

1. Few collections available
2. Focus on 2-class categorisation
3. No temporal dimension, no early risk analysis and detection
4. Little context about the user
5. Difficult to assess whether a mention of depression is genuine
1. large history for each redditor (several years)
2. many subreddits (communities) about different medical conditions (e.g. depression or anorexia)
3. long messages
4. terms & conditions of reddit allow us to use the data 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
building a redditor profile

retrieve all posting history
from any subreddit of
his/her posts +
his/her comments to other posts

often several years of text

removed of posts/comments with the explicit mention
of the diagnosis (depression group)
pre- & post-diagnosis text

organise the writings in chronological order into an XML archive

but removing the explicit mention of the diagnosis!
collections: main statistics

We created two collections for the early detection of depression

In 2017 and 2018 we focussed on early detection
early detection task

detect early traces of depression

for each subject, sequentially process all pieces of evidence...
early detection task

detect **early traces** of depression

for each subject, **sequentially process**

pieces of evidence...

John Doe's writings
(post or comments)

2/13/13
early detection task

detect early traces of depression

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

John Doe's writings (post or comments)

2/13/13
early detection task

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
early detection task

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  3/1/13
early detection task

detect early traces of depression

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

John Doe's writings (post or comments)

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2/13/13</td>
<td>2/15/13</td>
<td>3/1/13</td>
<td></td>
<td>12/9/16</td>
</tr>
</tbody>
</table>
early detection task

detect early traces of depression

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

tradeoff: early decision vs more informed decision

when should I fire an alarm?
performance metric

After seeing $k$ texts a system makes a binary decision $d$ about John Doe:

John Doe's writings (post or comments)

2/13/13 (1) 2/15/13 (2) ...

3/10/14 (k)

d=1 $\implies$ possible risk of depression
d=0 $\implies$ non-risk case
performance metric

John Doe's writings (post or comments)

2/13/13 (1)
2/15/13 (2)

3/10/14 (k)

decision (d)

Early Risk Detection Error:

\[ \text{ERDE}_0(d,k) = \begin{cases} 
\text{c}_{fp} & (\text{false positive}) \\
\text{c}_{fn} & (\text{false negative}) \\
\text{c}_{tp} \cdot \text{l}c_0(k) & (\text{true positive}) \\
0 & (\text{true negative}) 
\end{cases} \]

Usually, \( c_{fn} \gg c_{fp} \)

\( c_{fn} \leftarrow 1, c_{fp} \leftarrow \text{expected proportion of positive cases (e.g. 0.01)} \)
True Positive cost: $c_{tp} \times Ic_o(k)$

Latency cost function

$c_{tp} \leftarrow c_{fn}$ (late detection ≈ no detection)
Some results

Participants are getting better and better, but far from optimal!

<table>
<thead>
<tr>
<th></th>
<th>ERDE$_5$</th>
<th>ERDE$_{50}$</th>
<th>F1</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPLA</td>
<td>17.33%</td>
<td>15.83%</td>
<td>0.35</td>
<td>0.22</td>
<td>0.75</td>
</tr>
<tr>
<td>GPLB</td>
<td>19.14%</td>
<td>17.15%</td>
<td>0.30</td>
<td>0.18</td>
<td>0.83</td>
</tr>
<tr>
<td>GPLC</td>
<td>14.06%</td>
<td>12.14%</td>
<td>0.46</td>
<td>0.42</td>
<td>0.50</td>
</tr>
<tr>
<td>GPLD</td>
<td>14.52%</td>
<td>12.78%</td>
<td>0.47</td>
<td>0.39</td>
<td>0.60</td>
</tr>
<tr>
<td>FHDOA</td>
<td>12.82%</td>
<td>9.69%</td>
<td>0.64</td>
<td>0.61</td>
<td>0.67</td>
</tr>
<tr>
<td>FHDOB</td>
<td><strong>12.70%</strong></td>
<td><strong>10.39%</strong></td>
<td><strong>0.55</strong></td>
<td><strong>0.69</strong></td>
<td><strong>0.46</strong></td>
</tr>
<tr>
<td>FHDOE</td>
<td>13.24%</td>
<td>10.56%</td>
<td>0.56</td>
<td>0.57</td>
<td>0.56</td>
</tr>
<tr>
<td>FHDOD</td>
<td>13.04%</td>
<td>10.53%</td>
<td>0.57</td>
<td>0.63</td>
<td>0.52</td>
</tr>
<tr>
<td>UArizonaA</td>
<td>14.62%</td>
<td>12.68%</td>
<td>0.40</td>
<td>0.31</td>
<td>0.58</td>
</tr>
<tr>
<td>UArizonaB</td>
<td>13.07%</td>
<td>11.63%</td>
<td>0.30</td>
<td>0.33</td>
<td>0.27</td>
</tr>
<tr>
<td>UArizonaC</td>
<td>17.93%</td>
<td>12.74%</td>
<td>0.34</td>
<td>0.21</td>
<td><strong>0.92</strong></td>
</tr>
<tr>
<td>UArizonaD</td>
<td>14.73%</td>
<td>10.23%</td>
<td>0.45</td>
<td>0.32</td>
<td>0.79</td>
</tr>
<tr>
<td>UArizonaE</td>
<td>14.93%</td>
<td>12.01%</td>
<td>0.45</td>
<td>0.34</td>
<td>0.63</td>
</tr>
<tr>
<td>LyRA</td>
<td>15.65%</td>
<td>15.15%</td>
<td>0.14</td>
<td>0.11</td>
<td>0.19</td>
</tr>
<tr>
<td>LyRB</td>
<td>16.75%</td>
<td>15.76%</td>
<td>0.16</td>
<td>0.11</td>
<td>0.29</td>
</tr>
<tr>
<td>LyRC</td>
<td>16.14%</td>
<td>15.51%</td>
<td>0.16</td>
<td>0.12</td>
<td>0.25</td>
</tr>
<tr>
<td>LyRD</td>
<td>14.97%</td>
<td>14.47%</td>
<td>0.15</td>
<td>0.13</td>
<td>0.17</td>
</tr>
<tr>
<td>LyRE</td>
<td>13.74%</td>
<td>13.74%</td>
<td>0.08</td>
<td>0.11</td>
<td>0.06</td>
</tr>
<tr>
<td>UNSLA</td>
<td>13.66%</td>
<td><strong>9.68%</strong></td>
<td>0.59</td>
<td>0.48</td>
<td>0.79</td>
</tr>
<tr>
<td>UQAMA</td>
<td>14.03%</td>
<td>12.29%</td>
<td>0.53</td>
<td>0.48</td>
<td>0.60</td>
</tr>
<tr>
<td>UQAMB</td>
<td>13.78%</td>
<td>12.78%</td>
<td>0.48</td>
<td>0.49</td>
<td>0.46</td>
</tr>
<tr>
<td>UQAMC</td>
<td>13.58%</td>
<td>12.83%</td>
<td>0.42</td>
<td>0.50</td>
<td>0.37</td>
</tr>
<tr>
<td>UQAMD</td>
<td>13.23%</td>
<td>11.98%</td>
<td>0.38</td>
<td>0.64</td>
<td>0.27</td>
</tr>
<tr>
<td>UQAME</td>
<td>13.68%</td>
<td>12.68%</td>
<td>0.39</td>
<td>0.45</td>
<td>0.35</td>
</tr>
<tr>
<td>CHEPEA</td>
<td>14.75%</td>
<td>12.26%</td>
<td>0.48</td>
<td>0.38</td>
<td>0.65</td>
</tr>
<tr>
<td>CHEPEB</td>
<td>14.78%</td>
<td>12.29%</td>
<td>0.47</td>
<td>0.37</td>
<td>0.63</td>
</tr>
<tr>
<td>CHEPEC</td>
<td>14.81%</td>
<td>12.57%</td>
<td>0.46</td>
<td>0.37</td>
<td>0.63</td>
</tr>
<tr>
<td>CHEPED</td>
<td>14.81%</td>
<td>12.57%</td>
<td>0.45</td>
<td>0.36</td>
<td>0.62</td>
</tr>
<tr>
<td>NLPIA</td>
<td>15.59%</td>
<td>15.59%</td>
<td>0.15</td>
<td>0.12</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Table 3. Results
2019 & 2020: a new task

Different types of depression ...

Different levels of depression!

Solution: estimating the level of depression
Beck's Depression Inventory (BDI)

**BDI** is a 21-item, self-report rating inventory that measures characteristic, attitudes, and symptoms of depression.

The BDI takes approximately 10-15 minutes to complete.

Carried out as self-assessment or through a caregiver that ask the questions.
1. Sadness
0. I do not feel sad.
1. I feel sad much of the time.
2. I am sad all the time.
3. I am so sad or unhappy that I can't stand it.

2. Pessimism
0. I am not discouraged about my future.
1. I feel more discouraged about my future than I used to be.
2. I do not expect things to work out for me.
3. I feel my future is hopeless and will only get worse.

3. Past Failure
0. I do not feel like a failure.
1. I have failed more than I should have.
2. As I look back, I see a lot of failures.
3. I feel I am a total failure as a person.

4. Loss of Pleasure
0. I get as much pleasure as I ever did from the things I enjoy.
1. I don't enjoy things as much as I used to.
2. I get very little pleasure from the things I used to enjoy.
3. I can't get any pleasure from the things I used to enjoy.

5. Guilty Feelings
0. I don't feel particularly guilty.
1. I feel guilty over many things I have done or should have done.
2. I feel guilty about things I did not do.
3. I feel guilty about things I did.

   minimal depression (depression levels 0-9)
   mild depression (depression levels 10-18)
   moderate depression (depression levels 19-29)

6. Punishment
0. I don't feel severe depression (depression levels 30-63)
1. I feel I may be punished.
2. I expect to be punished.
3. I feel I am being punished.
Automatic filling of BDI using Text Mining

Ask eRsik participants to fill the questionnaire for the person the system is analyzing the posts

From manual to automatic filling
evaluation measures

- **Average Hit Rate (AHR):** hit rate averaged over all users
- **Average Closeness Rate (ACR):** closeness rate averaged over all users
- **Average Difference between Overall Depression Levels (ADODL):** difference between overall depression (DODL) averaged over all users
  - $DODL = (63 - ad_{overall})/63$
- **Depression Category Hit Rate (DCHR):** fraction of all cases where automated approach matches real questionnaire
Results from the 2019 eRisk

<table>
<thead>
<tr>
<th>Run</th>
<th>AHR</th>
<th>ACR</th>
<th>ADODL</th>
<th>DCHR</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioInfo@UAVR</td>
<td>34.05%</td>
<td>66.43%</td>
<td>77.70%</td>
<td>25.00%</td>
</tr>
<tr>
<td>BiTeM</td>
<td>32.14%</td>
<td>62.62%</td>
<td>72.62%</td>
<td>25.00%</td>
</tr>
<tr>
<td>CAMH_GPT_nearest_unsupervised</td>
<td>23.81%</td>
<td>57.06%</td>
<td>81.03%</td>
<td>45.00%</td>
</tr>
<tr>
<td>CAMH_GPT_supervised.181_features.58hr</td>
<td>35.47%</td>
<td>68.33%</td>
<td>75.63%</td>
<td>20.00%</td>
</tr>
<tr>
<td>CAMH_GPT_supervised.769_features.55lu</td>
<td>36.43%</td>
<td>67.22%</td>
<td>72.30%</td>
<td>20.00%</td>
</tr>
<tr>
<td>CAMH_GPT_supervised.949_features.75lu</td>
<td>36.91%</td>
<td>69.13%</td>
<td>75.63%</td>
<td>15.00%</td>
</tr>
<tr>
<td>CAMH_LIWC_supervised_SVM</td>
<td>35.95%</td>
<td>66.59%</td>
<td>75.48%</td>
<td>25.00%</td>
</tr>
<tr>
<td>Fazl</td>
<td>22.38%</td>
<td>56.27%</td>
<td>72.78%</td>
<td>5.00%</td>
</tr>
<tr>
<td>Illinois</td>
<td>22.62%</td>
<td>56.19%</td>
<td>66.35%</td>
<td>40.00%</td>
</tr>
<tr>
<td>ISIKol_multiSimilarity-5000-Dtac-Qtac</td>
<td>29.76%</td>
<td>57.94%</td>
<td>74.13%</td>
<td>25.00%</td>
</tr>
<tr>
<td>ISIKol-bm25-1.2-0.75-5000-Dtac-Qtac</td>
<td>29.76%</td>
<td>57.06%</td>
<td>72.78%</td>
<td>25.00%</td>
</tr>
<tr>
<td>ISIKol-lm-d-1.0-5000-Dtac-Qtac</td>
<td>30.00%</td>
<td>57.94%</td>
<td>73.02%</td>
<td>15.00%</td>
</tr>
<tr>
<td>Kimberly</td>
<td>38.33%</td>
<td>64.44%</td>
<td>66.19%</td>
<td>20.00%</td>
</tr>
<tr>
<td>UNSLA</td>
<td>37.38%</td>
<td>67.94%</td>
<td>72.86%</td>
<td>30.00%</td>
</tr>
<tr>
<td>UNSLB</td>
<td>36.93%</td>
<td>70.16%</td>
<td>76.83%</td>
<td>30.00%</td>
</tr>
<tr>
<td>UNSLC</td>
<td>41.43%</td>
<td>69.13%</td>
<td>78.02%</td>
<td>40.00%</td>
</tr>
<tr>
<td>UNSLD</td>
<td>38.10%</td>
<td>67.22%</td>
<td>78.02%</td>
<td>30.00%</td>
</tr>
<tr>
<td>UNSLE</td>
<td>40.71%</td>
<td>71.27%</td>
<td>80.48%</td>
<td>35.00%</td>
</tr>
<tr>
<td>all Os</td>
<td>34.52%</td>
<td>61.43%</td>
<td>61.43%</td>
<td>20.00%</td>
</tr>
<tr>
<td>all 1s</td>
<td>27.14%</td>
<td>71.75%</td>
<td>79.37%</td>
<td>20.00%</td>
</tr>
<tr>
<td>random (avg 1000 repetitions)</td>
<td>23.98%</td>
<td>58.55%</td>
<td>77.78%</td>
<td>33.55%</td>
</tr>
</tbody>
</table>

- **14 participants**
- **AHR**: best performing run got more than 40% answers right
- **ACR**: nobody beat the naïve all-1 algorithms
- **ADODL** and **DCHR**: best run get only 45% right answers

**Conclusions**: still far from an effective screening tool!
Results from the 2020 eRisk

Table 5. Task 3. Performance Results

<table>
<thead>
<tr>
<th>Run</th>
<th>AHR</th>
<th>ACR</th>
<th>ADODL</th>
<th>DCHR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biolinfo@UAVR</td>
<td>38.30%</td>
<td>69.21%</td>
<td>76.01%</td>
<td>30.00%</td>
</tr>
<tr>
<td>iLab run1</td>
<td>36.73%</td>
<td>68.68%</td>
<td>81.07%</td>
<td>27.14%</td>
</tr>
<tr>
<td>iLab run2</td>
<td>37.07%</td>
<td>69.41%</td>
<td>81.70%</td>
<td>27.14%</td>
</tr>
<tr>
<td>iLab run3</td>
<td>35.99%</td>
<td>69.14%</td>
<td>82.93%</td>
<td>34.29%</td>
</tr>
<tr>
<td>prhl_logreg_features</td>
<td>34.01%</td>
<td>67.07%</td>
<td>80.05%</td>
<td>35.71%</td>
</tr>
<tr>
<td>prhl_svm_use</td>
<td>36.94%</td>
<td>69.02%</td>
<td>81.72%</td>
<td>31.43%</td>
</tr>
<tr>
<td>prhl_svm_features</td>
<td>34.56%</td>
<td>67.44%</td>
<td>80.63%</td>
<td>35.71%</td>
</tr>
<tr>
<td>svm_features</td>
<td>34.56%</td>
<td>67.44%</td>
<td>80.63%</td>
<td>35.71%</td>
</tr>
<tr>
<td>relai_context_paral_user</td>
<td>36.80%</td>
<td>68.37%</td>
<td>80.84%</td>
<td>22.86%</td>
</tr>
<tr>
<td>relai_context_sim_answer</td>
<td>21.16%</td>
<td>55.40%</td>
<td>73.76%</td>
<td>27.14%</td>
</tr>
<tr>
<td>relai_lda_answer</td>
<td>28.50%</td>
<td>60.79%</td>
<td>79.07%</td>
<td>30.00%</td>
</tr>
<tr>
<td>relai_lda_user</td>
<td>36.39%</td>
<td>68.32%</td>
<td>83.15%</td>
<td>34.29%</td>
</tr>
<tr>
<td>relai_sylo_user</td>
<td>37.28%</td>
<td>68.37%</td>
<td>80.70%</td>
<td>20.00%</td>
</tr>
<tr>
<td>Run1_resulat_CNN_Methodode_max</td>
<td>34.97%</td>
<td>67.19%</td>
<td>76.85%</td>
<td>25.71%</td>
</tr>
<tr>
<td>Run2_resulat_CNN_Methodode_suite</td>
<td>32.79%</td>
<td>66.08%</td>
<td>76.33%</td>
<td>17.14%</td>
</tr>
<tr>
<td>Run3_resulat_BILSTM_Methodode_max</td>
<td>34.01%</td>
<td>67.78%</td>
<td>79.30%</td>
<td>22.86%</td>
</tr>
<tr>
<td>Run4_resulat_BILSTM_Methodode_suite</td>
<td>33.54%</td>
<td>67.26%</td>
<td>78.91%</td>
<td>20.00%</td>
</tr>
</tbody>
</table>

- More data, more speed required!
- Data (20 partic.) from 2019 was available
- Only 6 participants
- **AHR**: best performing run got only around 38% of answers right
  - Random alg. would give 25%
- **ADODL** a bit better, but **DCHR** much worse

Overall performance were even worse than 2019’s!
eRisk 2021: what are we planning?

• eRisk will run again with 3 tasks:
  1. Pathological gambling
  2. early detection of self-harm
  3. depression level estimation (DLE)

• For DLE:
  • Perform better analysis of which BDI questions are easier to answer
  • Propose to approach each question differently (i.e. a classifier for each question)
Conclusions

- new collections on depression & language use
- simple early risk detection algorithm (preliminary baselines)
- methodology for benchmark construction
- we focussed on the temporal dimension
Thank you for listening

Questions?
Acknowledgements

David E. Losada

Fabio Crestani

Swiss National Science Foundation
(project “Early risk prediction on the Internet: an evaluation corpus”, 2015)

“Ministerio de Economía y Competitividad”
of the Government of Spain &
FEDER Funds (ref. TIN2015-64282-R)