Multi-Aspect Transfer Learning for Detecting Low Resource Mental Disorders on Social Media

Ana-Sabina Uban, Berta Chulvi, Paolo Rosso
University of Bucharest, Romania
Universitat Politècnica de València, Spain

LREC
June 2022
Mental health

Depression

Depression (major depressive disorder) is a common and serious medical illness that negatively affects how you feel, the way you think and how you act.

Depression causes feelings of sadness and/or a loss of interest in activities you once enjoyed. It can lead to a variety of emotional and physical problems and can decrease your ability to function at work and at home.

Source: American Psychiatric Association website
Eating disorders

Eating disorders are illnesses in which the people experience severe disturbances in their eating behaviors and related thoughts and emotions. People with eating disorders typically become preoccupied with food and their body weight.

People with anorexia nervosa and bulimia nervosa tend to be perfectionists with low self-esteem and are extremely critical of themselves and their bodies.

Source: American Psychiatric Association website
Mental health

PTSD

Post-traumatic stress disorder (PTSD) is a psychiatric disorder that may occur in people who have experienced or witnessed a traumatic event such as a natural disaster, a serious accident, a terrorist act, war/combat, or rape or who have been threatened with death, sexual violence or serious injury.

People with PTSD have intense, disturbing thoughts and feelings related to their experience that last long after the traumatic event has ended. They may relive the event through flashbacks or nightmares; they may feel sadness, fear or anger; and they may feel detached or estranged from other people.

Source: American Psychiatric Association website
Mental health

Suicide prevention

As the 10th leading cause of death in the United States and the second leading cause of death (after accidents) for people aged 10 to 34, suicide is a serious public health problem.

Suicide is linked to mental disorders, particularly depression and alcohol use disorders.

Source: American Psychiatric Association website
Mental health disorders: Importance

- Affects quality of life (emotions, thoughts, activities, social)
- Affects physical health (sleep, eating, energy)
- Can lead to suicide
- COVID-19 pandemic affected mental health from multiple directions (health, social, economical, ...)
- Social media engagement can further affect mental health
- Underdiagnosed, undertreated
  - Depression 50% diagnosed, 13–49% properly treated
Mental disorders: automatic detection

Motivation and applicability

❖ **Alerting** users who show symptoms (recommend professional help); **suicide watch**, online counselling (chatbots) …
❖ Preventing development of disorders (**early** detection)
❖ **Assisting clinicians** with new insights: building, developing **diagnostic criteria** (e.g. anorexia)
  ➢ the diagnosis of certain disorders can also be a complicated issue, standards for diagnosis constantly evolving
  ➢ evidence of co-morbidity between certain disorders
Data for mental disorders

- Medical records
- Questionnaires
- Therapy sessions

16. Changes in Sleeping Pattern
   1. I have not experienced any change in my sleeping pattern.
   2a. I sleep somewhat more than usual.
   2b. I sleep somewhat less than usual.
   3a. I sleep a lot more than usual.
   3b. I sleep a lot less than usual.
   4a. I sleep most of the day.
   4b. I wake up 1-2 hours early and can’t get back to sleep.

17. Irritability
   1. I am no more irritable than usual.
   2. I am more irritable than usual.
   3. I am much more irritable than usual.
   4. I am irritable all the time.

18. Changes in Appetite
   1. I have not experienced any change in my appetite.
   2a. My appetite is somewhat less than usual.
   2b. My appetite is somewhat greater than usual.
   3a. My appetite is much less than before.
   3b. My appetite is much greater than usual.
   4a. I have no appetite at all.
   4b. I crave food all the time.

19. Concentration Difficulty
   1. I can concentrate as well as ever.
   2. I can’t concentrate as well as usual.
   3. It’s hard to keep my mind on anything for very long.
   4. I find I can’t concentrate on anything.

20. Tiredness or Fatigue
   1. I am no more tired or fatigued than usual.
   2. I get more tired or fatigued more easily than usual.
   3. I am too tired or fatigued to do a lot of the things I used to do.
   4. I am too tired or fatigued to do most of the things I used to do.

21. Loss of Interest in Sex
   1. I have not noticed any recent change in my interest in sex.
   2. I am less interested in sex than I used to be.
   3. I am much less interested in sex now.
   4. I have lost interest in sex completely.
Data for mental disorders

- Medical records
- Questionnaires
- Therapy sessions

Costly to annotate
Data for mental disorders

- Medical records
- Questionnaires
- Therapy sessions
- Social media

Table 1: Example titles of posts in the MHs and SW datasets; content has been carefully paraphrased to protect the privacy of the individuals.
Datasets for mental disorders

- Depression (mostly)
- Anorexia
- PTSD
- ...

Table 1: Example titles of posts in the MHs and SW datasets; content has been carefully paraphrased to protect the privacy of the individuals.
Research questions

**(RQ1)** Can transfer learning be leveraged in order to improve the detection performance of automatic deep learning models for disorders where datasets are scarce, and be used across different social media platforms?

**(RQ2)** What can we learn about the similarity between the different disorders through studying the effectiveness of transfer learning?

**(RQ3)** How can we use interpretable multi-aspect deep learning models to reveal qualitative conclusions about the specific linguistic dimensions which are more similar across different disorders?
Experimental setup

**Data:** social media posts collected based on self-stated diagnoses

Text classification: *supervised binary classification* at *user level* (is a user depressed...?); *cross-disorder* classification (what is this user suffering from...?)

**Deep** learning model, hierarchical architecture (post-level attention + user-level attention); *features* from multiple *levels* of the text: content, style and emotion features

**Transfer learning** experiments:
- Cross disorders
- Cross platform
- Comparing strategies
- Analyzing errors and useful features
Datasets

Workshops and shared tasks on mental disorder detection

**CLPsych**: Computational Linguistics and Clinical Psychology (2014, 2015,…)

- Linguistic Twitter data to detect various mental disorders

**eRisk**: Early Risk Detection on Social Media (since 2017)

- Textual data from reddit forums: depression, anorexia, self-harm…

Datasets used:

- depression (CLPsych, eRisk, + additional Twitter depression dataset)
- self-harm (eRisk)
- anorexia (eRisk)
- PTSD (CLPsych)

Annotated based on self-stated diagnoses
## Datasets statistics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Users</th>
<th>Positive %</th>
<th>Posts</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>eRisk self-harm <em>(reddit)</em></td>
<td>763</td>
<td>19%</td>
<td>274,534</td>
<td>~ 6M</td>
</tr>
<tr>
<td>eRisk anorexia <em>(reddit)</em></td>
<td>1287</td>
<td>10%</td>
<td>823,754</td>
<td>~ 23M</td>
</tr>
<tr>
<td>eRisk depression <em>(reddit)</em></td>
<td>1304</td>
<td>16%</td>
<td>811,586</td>
<td>~ 25M</td>
</tr>
<tr>
<td>CLPsych depression <em>(Twitter)</em></td>
<td>822</td>
<td>64%</td>
<td>1,919,353</td>
<td>~ 26M</td>
</tr>
<tr>
<td>CLPsych PTSD <em>(Twitter)</em></td>
<td>1078</td>
<td>72%</td>
<td>2,541,214</td>
<td>~ 19M</td>
</tr>
<tr>
<td>Twitter depression dataset</td>
<td>519</td>
<td>50%</td>
<td>52,080</td>
<td>~500K</td>
</tr>
</tbody>
</table>
Classification experiments:
Features

**Content:**
- Word sequences + word embeddings (GloVe)

**Style:**
- Function words (as bag of words)

**Emotion:**
- NRC emotion lexicon (as proportion of each emotion in each post)

**LIWC** categories (topics, emotions, style) (as proportion of each category in each post)
Classification experiments
Features

NRC emotions (Plutchik’s 8 emotions + 2 sentiments):

anger, anticipation, disgust, fear, joy, sadness, surprise, trust; negative, positive

LIWC categories (64 categories):

➢ Sentiment polarity
➢ Emotions (sadness, anxiety, affect...)
➢ Syntactic categories (pronouns, verbs, conjunctions...)
➢ Topics (health, money, religion, work...)
Our solution: model architecture
Classification results: cross-disorder classification

Depression vs self-harm vs anorexia classification (Reddit): **0.44 F1**
Depression vs PTSD classification (Twitter): **0.72 F1**

### Reddit

<table>
<thead>
<tr>
<th></th>
<th>Depr</th>
<th>Self-harm</th>
<th>Anorexia</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Depr</td>
<td>139</td>
<td>2</td>
<td>113</td>
</tr>
<tr>
<td>True Self-harm</td>
<td>60</td>
<td>67</td>
<td>144</td>
</tr>
<tr>
<td>True Anorexia</td>
<td>201</td>
<td>16</td>
<td>218</td>
</tr>
</tbody>
</table>

### Twitter

<table>
<thead>
<tr>
<th></th>
<th>Depr</th>
<th>PTSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Depr</td>
<td>126</td>
<td>24</td>
</tr>
<tr>
<td>True PTSD</td>
<td>65</td>
<td>95</td>
</tr>
</tbody>
</table>

Confusion matrices for classification between disorders
Transfer learning

Strategy 0. Zero-shot

Depression data
Transfer learning

**Strategy 1. Transfer layer**

Example: cross-task (depression → anorexia)
Transfer learning

**Strategy 2. Fine-tuning**

Example: cross-task (depression → anorexia)
Transfer learning

**Strategy 3. Multi-task learning**

Example: cross-platform (reddit / Twitter)
Transfer learning experiments.

Results

<table>
<thead>
<tr>
<th>Source</th>
<th>CROSS-DISORDER</th>
<th>CROSS-PLATFORM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>eRisk depression</td>
<td>eRisk depression</td>
</tr>
<tr>
<td></td>
<td>CLPsych depression</td>
<td>CLPsych depression</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>AUC</td>
</tr>
<tr>
<td>Strategy 0</td>
<td>.17</td>
<td>.62</td>
</tr>
<tr>
<td>Strategy 1</td>
<td>.64</td>
<td>.90</td>
</tr>
<tr>
<td>Strategy 2</td>
<td>.63</td>
<td>.93</td>
</tr>
<tr>
<td>Baseline HAN</td>
<td>.46</td>
<td>.91</td>
</tr>
</tbody>
</table>

Cross-disorder and cross-platform transfer learning results, compared to individual disorder prediction

<table>
<thead>
<tr>
<th>Source</th>
<th>All depression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>eRisk</td>
</tr>
<tr>
<td></td>
<td>F1</td>
</tr>
<tr>
<td>Strategy 3</td>
<td>.39</td>
</tr>
<tr>
<td>Single-task</td>
<td>.44</td>
</tr>
</tbody>
</table>

Cross-platform multi-task learning results
Transfer learning experiments.

Ablation results for cross-disorder transfer learning experiments (fine-tuning strategy)

<table>
<thead>
<tr>
<th>Source Target</th>
<th>eRisk</th>
<th>CLPsych</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Anorexia</td>
<td>Self-harm</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>AUC</td>
</tr>
<tr>
<td>All-word seq</td>
<td>.49</td>
<td>.88</td>
</tr>
<tr>
<td>All-function words</td>
<td>.51</td>
<td>.90</td>
</tr>
<tr>
<td>All-lexicon feat</td>
<td>.50</td>
<td>.91</td>
</tr>
<tr>
<td>All features</td>
<td>.63</td>
<td><strong>.93</strong></td>
</tr>
</tbody>
</table>
Transfer learning experiments.
Error analysis

Features with highest differences between correctly classified and misclassified texts.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Psycho-linguistic categories (LIWC features)</th>
<th>Emotions (NRC features)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression</td>
<td>verbs, tentative, I (1st pers pron), adverbs, past tense, pronouns, present tense, conjunctions</td>
<td>fear, anger, negative emotion, sadness</td>
</tr>
<tr>
<td>eRisk baseline</td>
<td>health, insight, cognitive processes, pronouns function words, adverbs</td>
<td>sadness, negative emotion</td>
</tr>
<tr>
<td>Self-harm</td>
<td>future tense, positive emotion, affective, function words, adverbs, present tense, pronouns</td>
<td>anger, fear, negative emotion</td>
</tr>
<tr>
<td>baseline</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anorexia</td>
<td>they (3rd pers pron), health, insight, she/he</td>
<td>fear, joy, positive emotion, negative emotion, sadness</td>
</tr>
<tr>
<td>baseline</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PTSD</td>
<td>you (2nd pers pron), function words, impersonal pronouns, verbs</td>
<td>positive emotion</td>
</tr>
<tr>
<td>baseline</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depr $\rightarrow$ self-harm transfer</td>
<td>future tense, affective, function words, adverbs, present tense, I (1st pers pron), verbs, social</td>
<td>fear, negative emotion</td>
</tr>
<tr>
<td>Depr $\rightarrow$ anorexia transfer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depr $\rightarrow$ PTSD transfer</td>
<td>exclusive, sad, conjunctions, adverbs, friend, biology</td>
<td>anger, positive emotion, sadness</td>
</tr>
</tbody>
</table>
Conclusions & future work

Our experiments have shown that transfer learning could be leveraged to build detection models for disorders where annotated data is scarce.

We have investigated and demonstrated the similarity between manifestations of different disorders at different levels of language (some more than others).

**Future**: multi-modal solutions and sentence embeddings as models; additional disorders with known comorbidities.
Thank you!
¡Gracias!
Mersi!
Merci!