Deep learning for detecting mental health disorders in social media text

Ana-Sabina Uban
(+Paolo Rosso, Berta Chulvi)

BioMedical NLP

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

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

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

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

Motivation

- 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

Applications

- **Alerting** users who show symptoms (recommend professional **help**)
- **Suicide watch**, online counselling (chatbots) …
- Preventing development of disorders (**early** detection)
- **Assisting clinicians** with new insights and building **diagnosis tools** (patterns of depressive symptoms, causes of depression,…)

Data for mental disorders

- Medical records
- Questionnaires
- Therapy sessions
- Essays, letters etc

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

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

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

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

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

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

Data for mental disorders

- Medical records
- Questionnaires
- Therapy sessions
- Essays, letters etc
- Social media

MHs (Mental Health subreddits)
I have been considering going for some formal therapy. Any suggestions?
Everyday I feel sad and lonely
Since past sometime I think I am having panic attacks. I really need help from you guys.
It has been so many years, I feel I still can't move on. I am noticing behavior what could be considered "triggers" now.
SW (SuicideWatch)
I know I was never meant to lead this life.
Don't want to hurt the people I care but I can't take this anymore.
Today I felt I have nothing left, why am I even living... I don't see a point.
I'd kill myself, but the other part of me tells me not to waste all the money my parents invested on me..

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 - mental illness in social media

Types of assessment - establishing ground truth:

- Annotated data
  - Collecting public posts of users selected from medical records / who answered questionnaires
- Self-stated diagnosis
  - Users who have shared their mental health diagnosis (identified through keyword searches: "`I have been diagnosed with depression`")
  - Users active on mental illness related forums (e.g. /r/depression, /r/anxiety, ...)

Research: Workshops and shared tasks

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

- ❖ Linguistic Twitter data to detect various mental disorders

AVAC: Audio-Video Affect Challenge (since 2010)

- ❖ Video, audio, text interviews; interview-level labels (The Distress Analysis Interview Corpus of human and computer interviews)
- ❖ Task: predict severity of depression
- ❖ Various adjacent shared tasks (cross-cultural affect etc)

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

- ❖ Textual data from reddit forums
  - ➢ Depression (+severity)
  - ➢ Anorexia
  - ➢ Self-harm

Previous approaches - Results

How difficult is mental disorder detection?

“Social media-based screening may reach prediction performance somewhere between unaided clinician assessment and screening surveys.” ([Detecting depression and mental illness on social media: an integrative review](#))

AUC moderate to high (0.6-0.9 AUC)

Early detection: more challenging (0.65-0.75 F1)

- Harder to detect before the onset of the mental illness

Mental disorder detection
Previous approaches

Features:

- Lexicons: LIWC (self-references, social words, emotion words, cognitive words.)
- Character n-grams, bag-of-words
- Topic modelling (sentiment-bearing topics, topic model with depression seed words, ...)
- Meta: user activity (social engagement, login times), demographic attributes (gender, age)
- Multimodal (rare): video interviews, profile picture
- Recently: language models (contextual embeddings, neural language models)

Models:

- SVM, random forest, neural networks
- Last couple of years: hierarchical attention networks, transformers

Features correlated with depression

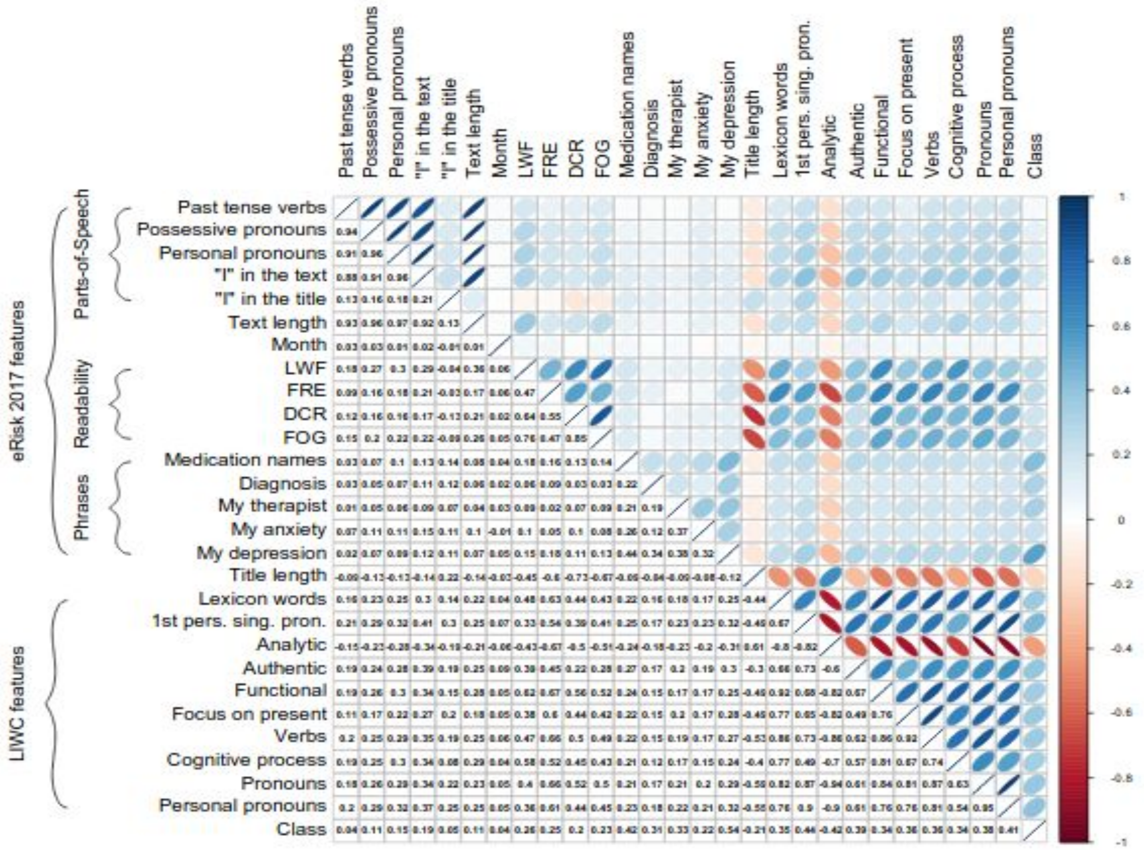


Fig. 1. Correlation matrix of all user features including the class information (non-depressed/depressed) based on the depression subtask training data. This plot is best viewed in electronic form.

[Word Embeddings and Linguistic Metadata atthe CLEF 2018 Tasks for Early Detection ofDepression and Anorexia]()

One solution:
mental disorder detection with deep learning

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

Text classification: **supervised binary classification** at **user level** (is a user depressed...?)

Deep learning model (neural networks): LSTM + attention

Hierarchical architecture (post-level attention + user-level attention)

Features from multiple **levels** of the text: content, style and emotion features

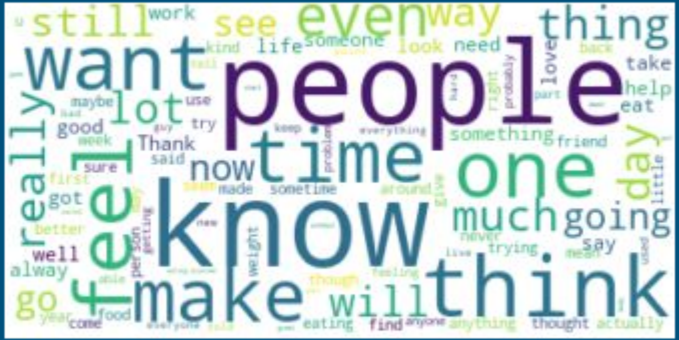
Interpretability

Datasets

Reddit (eRisk workshop)

DEPRESSION

ANOREXIA



Twitter (CLPsych workshop)

SELF-HARM

PTSD

Datasets statistics

Dataset	Users	Positive %	Posts	Words
eRisk self-harm (reddit)	763	19%	274,534	~ 6M
eRisk anorexia (reddit)	1287	10%	823,754	~ 23M
eRisk depression (reddit)	1304	16%	811,586	~ 25M
CLPsych depression (Twitter)	822	64%	1,919,353	~ 26M
CLPsych PTSD (Twitter)	1078	72%	2,541,214	~ 19M

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*…)

Preprocessing & feature extraction

Input tweet (user with PTSD)

'RT @1OR42wsOEFcv3f: I fall to fast, crash too hard, forgive too easily and care too much.'

Normalize & tokenize (20,000 words vocab)

i fall fast crash too hard forgive too easily and care too much

Extract features:

Stopwords

i fall fast too hard forgive too and care too much → [1, 0, 0, 2, 1, 0, 0, ..]

Emotions & categories

pronoun feel affect

i fall fast too hard forgive too and care too much → [0.06, 0, 0.02, 0.1, ..]

Word embeddings sequences

i fall fast too hard forgive too and care too much →

0.16, 0.22, 0.92, 0.12
0.07, 0.04, 0.22, 0.94
0.82, 0.11, 0.42, 0.81
0.06, 0.53, 0.02, 0.41

Encoding texts

Generating datapoints

User-level label (1 / 0)

Chunks of 50 posts

	text	subject	label
	@x3Qk4teUohz_ @naQ0WGvAGW all guns bought frim ffl dealers already have BGC and all shops that sell guns have to be ffl dealers so NO	eNBwLZDkkE	1
	@phm53IYapYEHKp @mESTieZqJN5m7K I prefer it myself but then I have a vest that I used to wear horseback cross draw was more comfortable	eNBwLZDkkE	1
	@uz69PsBVIERg @caND7HgdWcB1 @sQwDFKH5n72h @poWr6B1 @okj8UBit3Av I didn't know I had that much ammo? Did you send me a bday gift? Lol	eNBwLZDkkE	1
	@naQ0WGvAGW @ihQfDgubNLxrbHN just one of many reasons she can't win in Texas we don't have any bcg loopholes idiot	eNBwLZDkkE	1
	@oNz3gba2 When in fact you can't do anything to prevent someone from buying a gun that has not yet committed a crime	eNBwLZDkkE	1
	...	...	...
	RT @vLCl7uvpccHUfff: We should all thrive to be this ..well if you're a dog owner http://t.co/lWhwcWRpAE	eNBwLZDkkE	1
	@oNmEfFcOfMopi @dZf_sFui1dJ @mMne7kONGC @wtTIz9KuIOzRM @sfUnf28D inversion table yes not gravity boots can't get down without help	eNBwLZDkkE	1
	RT @hMHx8VCkRuK3: Attention Politicians!! #WeThePeople Own This Country. U WORK FOR US!! #RedNationRising http://t.co/j_hc0K7KCq v @wKe14R3…	eNBwLZDkkE	1
	@wSI0FaZC @bTUS3xYBh @h9_00Ot_dR4bFe @q0FRoB9wbWZRH well on obamacare alone he has changed law and delayed parts authority he does not have	eNBwLZDkkE	1
	@wSI0FaZC Don't have the final total yet still waiting on the IRS to notify me about how much my fine will be	eNBwLZDkkE	1

Posts truncated/padded to 256 words

Experimental setup

- Training / validation / test split:
  - Preserving train/test split in original paper
  - Training and test data are **disjoint** at user level
- Classification of individual posts poor => **chunking** posts (1 datapoint = 50 concatenated posts from 1 user)
- Data imbalance => **weighted loss**
- **Regularization:** dropout
- Batch normalization (before concatenation of different features)
- Adam optimizer

Hierarchical Attention Network

(Hierarchical Attention Networks for Document Classification)

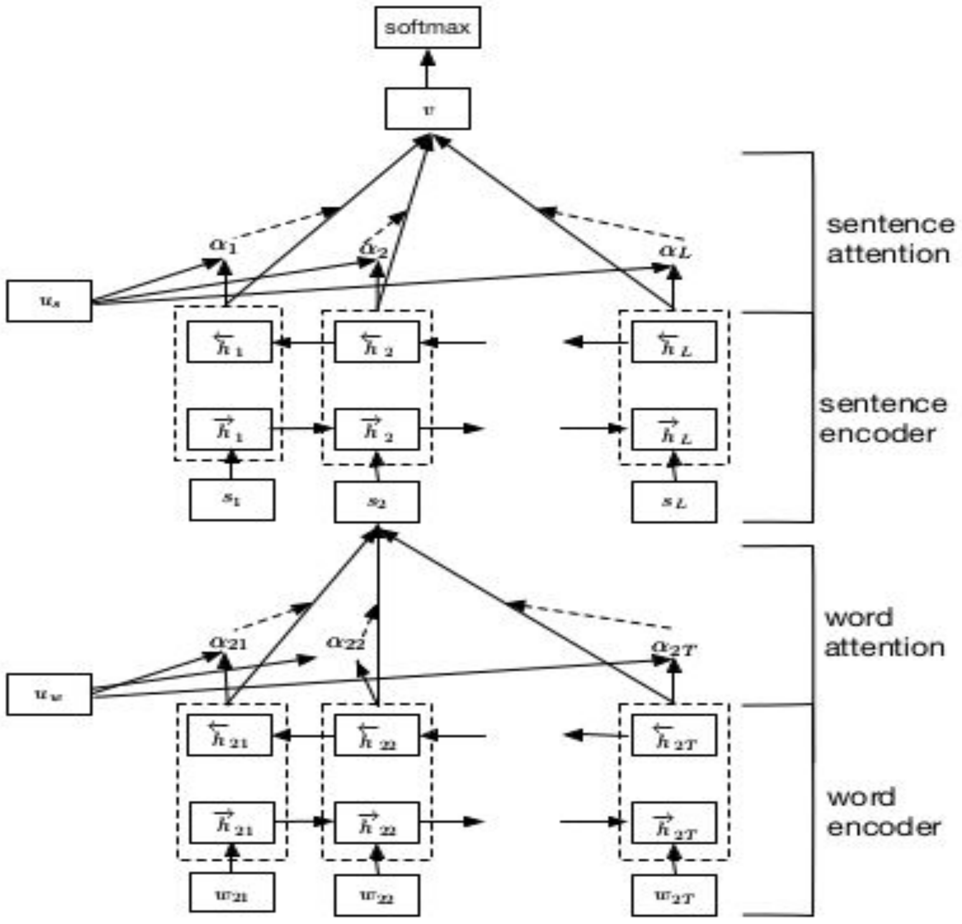
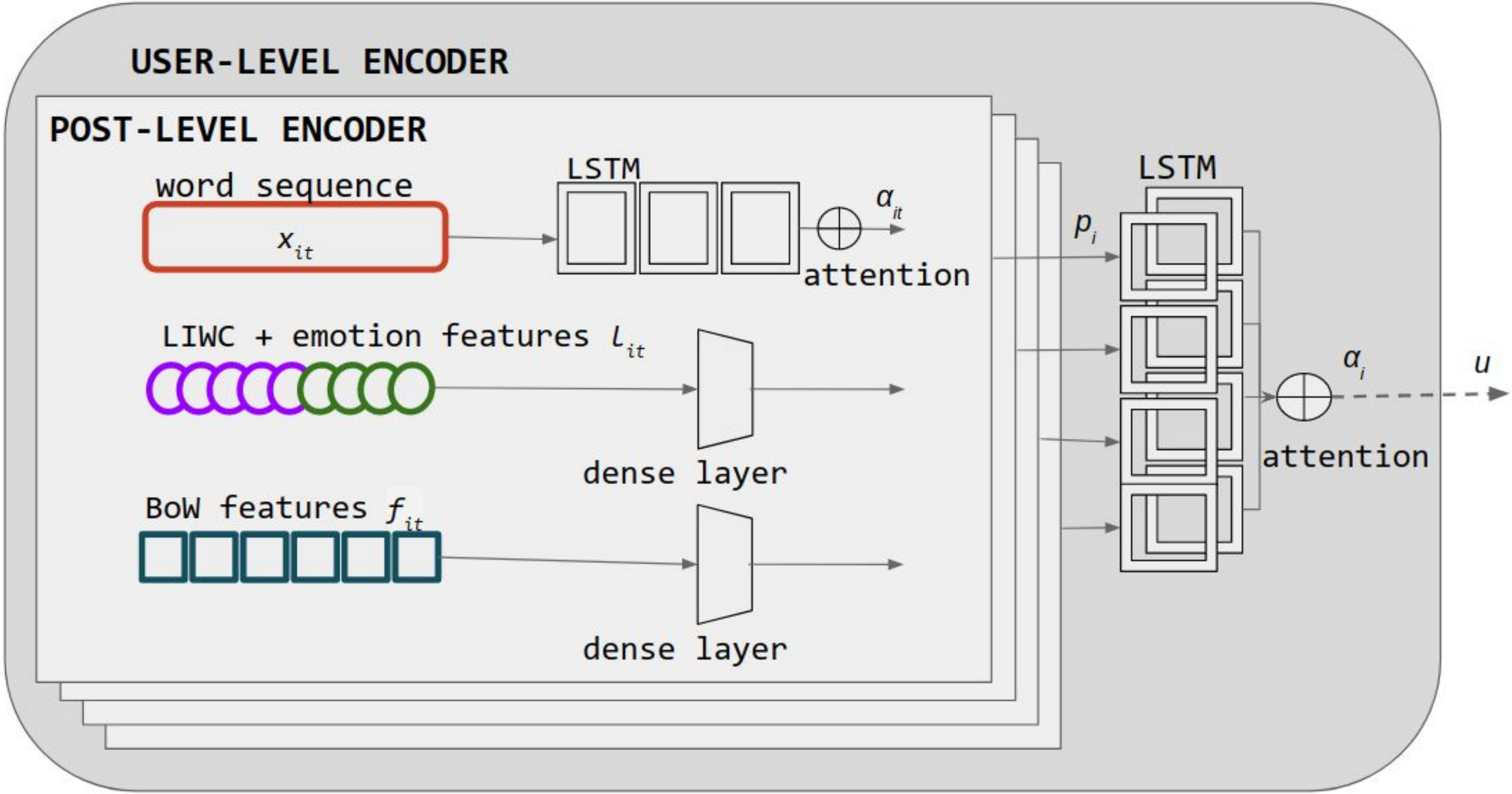


Figure 2: Hierarchical Attention Network.

Our solution: model architecture



Hierarchical attention

Post-level encoder

$$x_{it} = W_e w_{it}, t \in [1, T]$$
$$\overrightarrow{h}_{it} = \overrightarrow{f}(x_{it}), t \in [1, T]$$
$$v_{it} = tanh(W_w h_{it} + b_w)$$
$$\alpha_{it} = \frac{exp(v_{it}^T v_w)}{\sum_t exp(v_{it}^T v_w)}$$
$$s_i = \sum_t \alpha_{it} h_{it}$$

WORD SEQs

$$hf_{it} = W_f f_i + b_f$$

STOPWORDS

$$hl_{it} = W_l l_i + b_l$$

LEXICON

$$p_i = s_i \oplus hf_{it} \oplus hl_{it}$$

USER ENCODING

User-level encoder

$$h_i = \overrightarrow{LSTM}(p_i)$$
$$v_i = tanh(W_p h_i + b_p)$$
$$\alpha_i = \frac{exp(v_i^T v_p)}{\sum_t exp(v_i^T v_p)}$$
$$u = \sum_i \alpha_i h_i$$

Post encoder
(word level)

post-level attention

Layer (type)	Output Shape	Param #	Connected to
word_seq (InputLayer)	[(None, 256)]	0	
embeddings_layer (Embedding)	(None, 256, 100)	2000200	word_seq[0][0]
embedding_dropout (Dropout)	(None, 256, 100)	0	embeddings_layer[0][0]
LSTM_layer (LSTM)	(None, 256, 128)	117248	embedding_dropout[0][0]
attention (Dense)	(None, 256, 1)	129	LSTM_layer[0][0]
flatten (Flatten)	(None, 256)	0	attention[0][0]
activation (Activation)	(None, 256)	0	flatten[0][0]
repeat_vector (RepeatVector)	(None, 128, 256)	0	activation[0][0]
permute (Permute)	(None, 256, 128)	0	repeat_vector[0][0]
multiply (Multiply)	(None, 256, 128)	0	LSTM_layer[0][0] permute[0][0]
lambda (Lambda)	(None, 128)	0	multiply[0][0]
sent_repr_dropout (Dropout)	(None, 128)	0	lambda[0][0]

Total params: 2,117,577
Trainable params: 2,117,577
Non-trainable params: 0

User encoder (full)

WORD SEQs

LEXICON

STOPWORDS

USER ENCODING

user-level attention

Layer (type)	Output Shape	Param #	Connected to
hierarchical_word_seq_input (In	[(None, 50, 256)]	0	
numeric_input_hist (InputLayer)	[(None, 50, 75)]	0	
sparse_input_hist (InputLayer)	[(None, 50, 179)]	0	
post encoder (TimeDistributed)	(None, 50, 128)	2117577	hierarchical_word_seq_input[0][0]
numerical_dense_layer_user (Tim	(None, 50, 20)	1520	numeric_input_hist[0][0]
sparse_dense_layer_user (TimeDi	(None, 50, 20)	3600	sparse_input_hist[0][0]
concatenate (Concatenate)	(None, 50, 168)	0	user_encoder[0][0] numerical_dense_layer_user[0][0] sparse_dense_layer_user[0][0]
LSTM_layer_user (LSTM)	(None, 50, 32)	25728	concatenate[0][0]
attention_user (Dense)	(None, 50, 1)	33	LSTM_layer_user[0][0]
flatten_1 (Flatten)	(None, 50)	0	attention_user[0][0]
activation_1 (Activation)	(None, 50)	0	flatten_1[0][0]
repeat_vector_1 (RepeatVector)	(None, 32, 50)	0	activation_1[0][0]
permute_1 (Permute)	(None, 50, 32)	0	repeat_vector_1[0][0]
multiply_1 (Multiply)	(None, 50, 32)	0	LSTM_layer_user[0][0] permute_1[0][0]
lambda_1 (Lambda)	(None, 32)	0	multiply_1[0][0]
user_repr_dropout (Dropout)	(None, 32)	0	lambda_1[0][0]
output_layer (Dense)	(None, 1)	33	user_repr_dropout[0][0]

Total params: 2,148,491
Trainable params: 2,148,491
Non-trainable params: 0

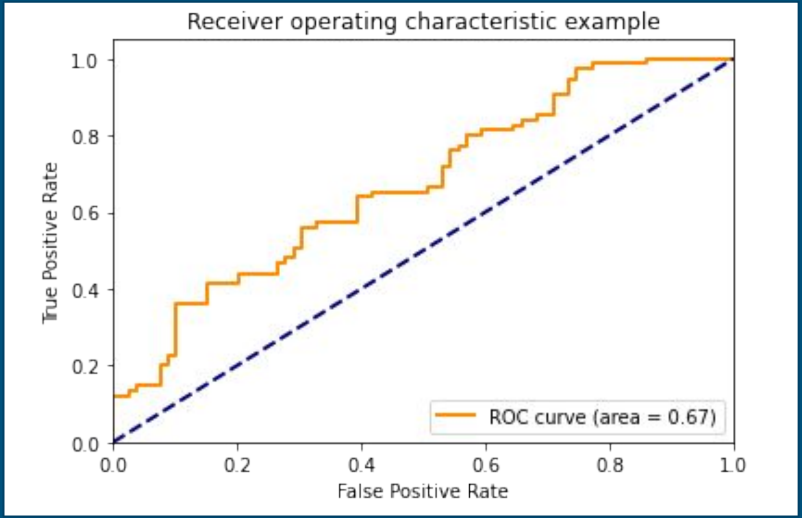
Attention implementation

```python
# Attention
if 'attention' not in ignore_layer:
    attention_layer = Dense(1, activation='tanh', name='attention')
    attention = attention_layer(lstm_layers)
    attention = Flatten()(attention)
    attention_output = Activation('softmax')(attention)
    attention = RepeatVector(hyperparams['lstm_units'])(attention_output)
    attention = Permute([2, 1])(attention)

    sent_representation = Multiply()([lstm_layers, attention])
    sent_representation = Lambda(lambda xin: K.sum(xin, axis=1),
                                 output_shape=(hyperparams['lstm_units'],)
                                )(sent_representation)
```

Evaluation

- Evaluation Metrics
  - Precision, recall, F1-score (positive class)
  - AUC (ROC) score <- data imbalance

- Baseline model
  - Logistic regression, transformers
  - with bag of word features



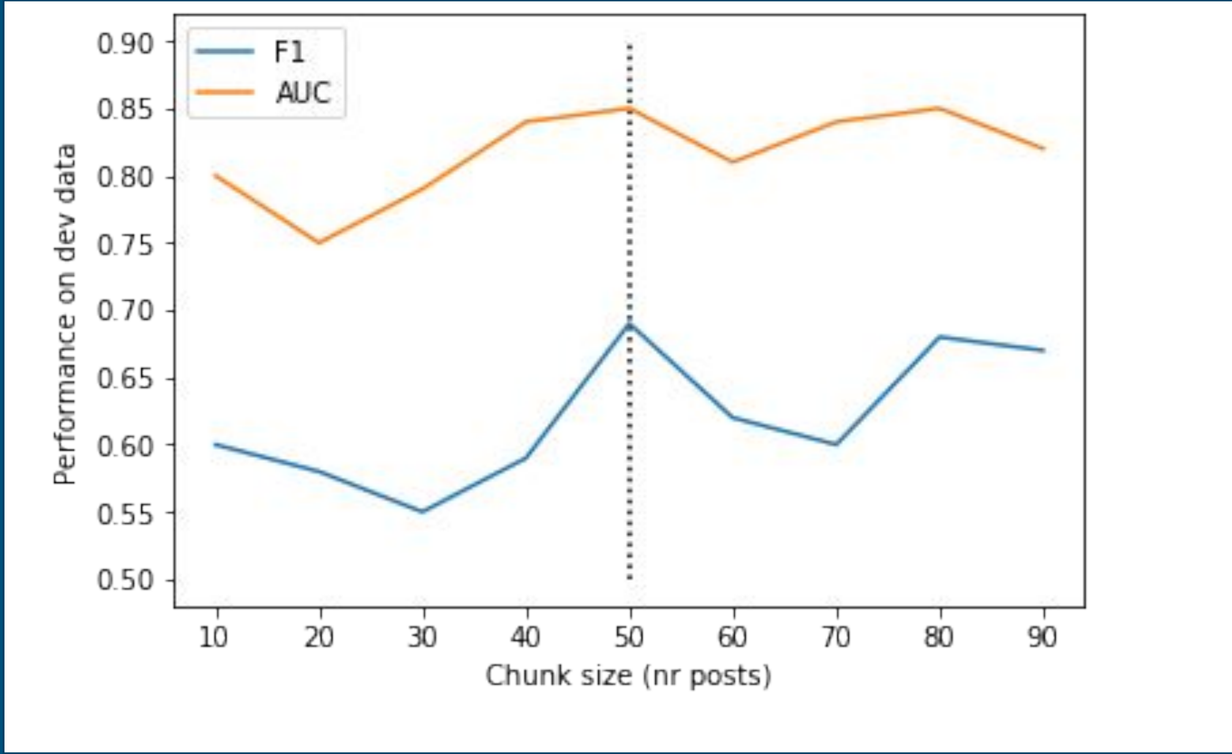
Results

	Depression (reddit)		Anorexia (reddit)		Self-harm (reddit)		Depression (Twitter)		PTSD (Twitter)	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
BiLSTM	.40	.82	.53	.90	.62	.84	.56	.72	.55	.78
CNN+LSTM	.35	.80	.76	.95	.44	.82	.56	.72	.61	.77
HAN	.44	.85	.61	.96	.65	.87	.53	.73	.57	.70
LogReg	.36	.76	.49	.90	.45	.75	.55	.72	.49	.69
RoBERTa	.40	.71	.70	.83	.35	.60	.54	.65	.40	.57

Findings

- Dataset size important (more data => better performance with DL)
- Better results on reddit than Twitter datasets
- Freezing vs training embedding weights for our task: training the embeddings gives better results (domain adaptation?)
- Bigger chunks (more text in 1 datapoint) help with performance

Performance ~ number of posts
Self-harm detection



Performance ~ number of posts

Anorexia detection

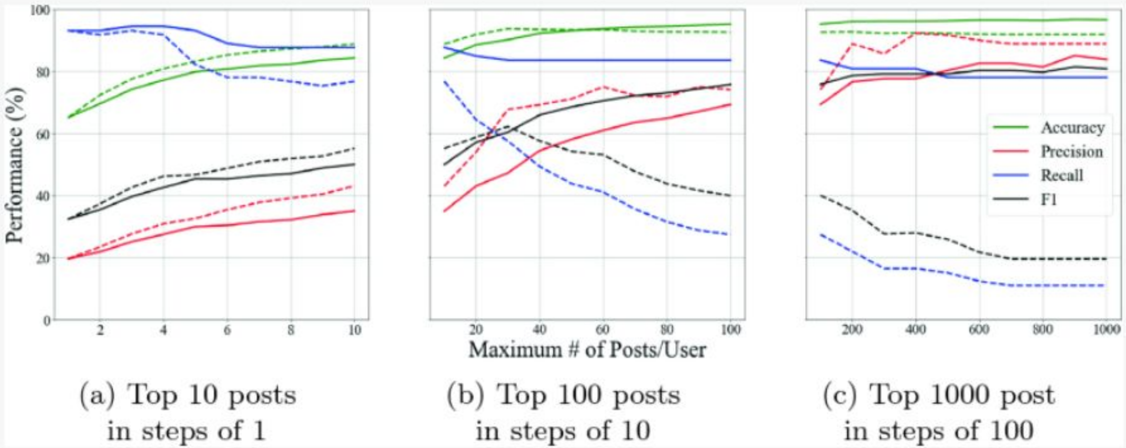


Fig. 2.

Performance of the system in terms of the maximum number of highly-weighted posts from each user. The solid lines correspond to the model with user-level attention (experiment 1), while the dotted lines correspond to the model with user-level average pooling (experiment 2).

Amini, Hessam, and Leila Kosseim. "Towards Explainability in Using Deep Learning for the Detection of Anorexia in Social Media." In International Conference on Applications of Natural Language to Information Systems, pp. 225-235. Springer, Cham, 2020.

Explainability

Neural networks are powerful models, but often act as **black boxes**.

Impediment for building **applications:** applications in medical domain can have serious impact on people's lives => need **trust** in models.

Regulations for **interpretability** of models in medical/mental health domain (e.g. GDPR in EU). [Current Regulation of Mobile Mental Health Applications]

Techniques:

- ➢ Attention weights
- ➢ Ablation experiments
- ➢ Error analysis
- ➢ Feature-level analysis
- ➢ Hidden layer activations/weights analysis

Ablation

WORD SEQs

LEXICON

STOPWORDS

USER ENCODING

user-level attention

```
Layer (type)                     Output Shape         Param #     Connected to
==================================================================================================
hierarchical_word_seq_input (In  [(None, 50, 256)]    0
__________________________________________________________________________________________________
numeric_input_hist (InputLayer)  [(None, 50, 75)]     0
__________________________________________________________________________________________________
sparse_input_hist (InputLayer)   [(None, 50, 179)]    0
__________________________________________________________________________________________________
post encoder (TimeDistributed)   (None, 50, 128)      2117577     hierarchical_word_seq_input[0][0]
__________________________________________________________________________________________________
numerical_dense_layer_user (Tim  (None, 50, 20)       1520        numeric_input_hist[0][0]
__________________________________________________________________________________________________
sparse_dense_layer_user (TimeDi  (None, 50, 20)       3600        sparse_input_hist[0][0]
__________________________________________________________________________________________________
concatenate (Concatenate)        (None, 50, 168)      0           user_encoder[0][0]
                                                                  numerical_dense_layer_user[0][0]
                                                                  sparse_dense_layer_user[0][0]
__________________________________________________________________________________________________
LSTM_layer_user (LSTM)           (None, 50, 32)       25728       concatenate[0][0]
__________________________________________________________________________________________________
attention_user (Dense)           (None, 50, 1)        33          LSTM_layer_user[0][0]
__________________________________________________________________________________________________
flatten_1 (Flatten)              (None, 50)           0           attention_user[0][0]
__________________________________________________________________________________________________
activation_1 (Activation)        (None, 50)           0           flatten_1[0][0]
__________________________________________________________________________________________________
repeat_vector_1 (RepeatVector)   (None, 32, 50)       0           activation_1[0][0]
__________________________________________________________________________________________________
permute_1 (Permute)              (None, 50, 32)       0           repeat_vector_1[0][0]
__________________________________________________________________________________________________
multiply_1 (Multiply)            (None, 50, 32)       0           LSTM_layer_user[0][0]
                                                                  permute_1[0][0]
__________________________________________________________________________________________________
lambda_1 (Lambda)                (None, 32)           0           multiply_1[0][0]
__________________________________________________________________________________________________
user_repr_dropout (Dropout)      (None, 32)           0           lambda_1[0][0]
__________________________________________________________________________________________________
output_layer (Dense)             (None, 1)            33          user_repr_dropout[0][0]
==================================================================================================
Total params: 2,148,491
Trainable params: 2,148,491
Non-trainable params: 0
```

Ablation

WORD SEQs

LEXICON

STOPWORDS

USER ENCODING

user-level attention

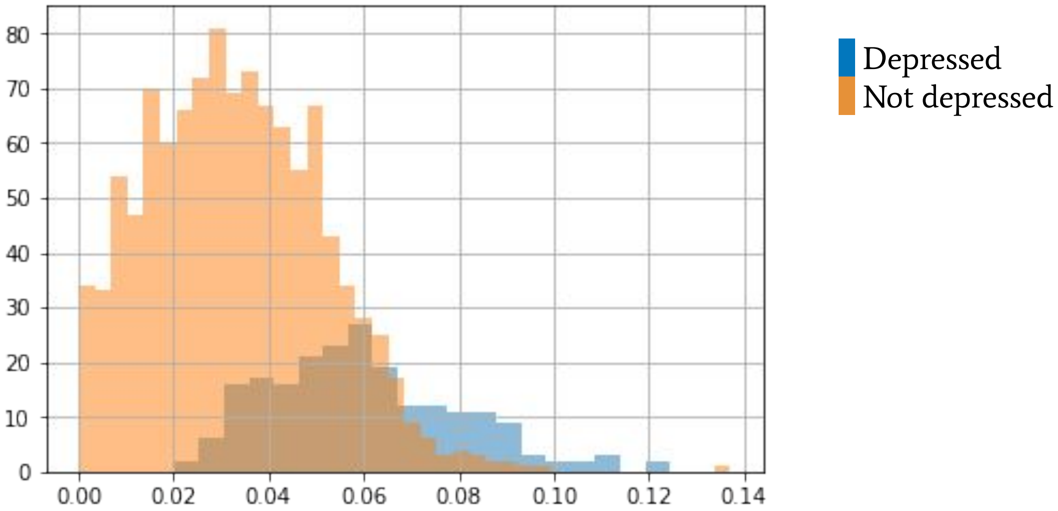
```
Layer (type)                    Output Shape        Param #     Connected to
==================================================================================================
hierarchical_word_seq_input (In [(None, 50, 256)]   0

numeric_input_hist (InputLayer) [(None, 50, 75)]    0

sparse_input_hist (InputLayer)  [(None, 50, 179)]   0

post encoder (TimeDistributed)  (None, 50, 128)     2117577     hierarchical_word_seq_input[0][0]

numerical_dense_layer_user (Tim (None, 50, 20)      1520        numeric_input_hist[0][0]

sparse_dense_layer_user (TimeDi (None, 50, 20)      3600        sparse_input_hist[0][0]

concatenate (Concatenate)       (None, 50, 168)     0           user_encoder[0][0]
                                                                numerical_dense_layer_user[0][0]
                                                                sparse_dense_layer_user[0][0]

LSTM_layer_user (LSTM)          (None, 50, 32)      25728       concatenate[0][0]

attention_user (Dense)          (None, 50, 1)       33          LSTM_layer_user[0][0]

flatten_1 (Flatten)             (None, 50)          0           attention_user[0][0]

activation_1 (Activation)       (None, 50)          0           flatten_1[0][0]

repeat_vector_1 (RepeatVector)  (None, 32, 50)      0           activation_1[0][0]

permute_1 (Permute)             (None, 50, 32)      0           repeat_vector_1[0][0]

multiply_1 (Multiply)           (None, 50, 32)      0           LSTM_layer_user[0][0]
                                                                permute_1[0][0]

lambda_1 (Lambda)               (None, 32)          0           multiply_1[0][0]

user_repr_dropout (Dropout)     (None, 32)          0           lambda_1[0][0]

output_layer (Dense)            (None, 1)           33          user_repr_dropout[0][0]
==================================================================================================
Total params: 2,148,491
Trainable params: 2,148,491
Non-trainable params: 0
```

Ablation results

	Depression (reddit)		Anorexia (reddit)		Self-harm (reddit)		Depression (Twitter)		PTSD (Twitter)	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
HAN	.44	.85	.61	.96	.65	.87	.53	.73	.57	.70
HAN-word sequences	.33	.81	.48	.91	.34	.83	.51	.68	.52	.65
HAN-stopwords	.55	.84	.47	.95	.55	.84	.53	.69	.56	.69
HAN-emotion features	.37	.85	.45	.94	.59	.86	.52	.68	.52	.68
HAN-LIWC features	.43	.84	.45	.91	.62	.87	.50	.67	.54	.68

Feature analysis - “I”

Depression



The use of “I” in depressed vs non-depressed users

Feature analysis: category-label correlations

Depression

health, certain, feel, you, negate, social, tentative, future, cognitive processes, present, conjunction, pronoun, function words, verb, future, I, work, leisure, money, space, death, fear,

Self-harm

negemo, past, sadness, health, adverb, present, future, cognitive processes, pronoun, function words, I, work, we, leisure, positive,

Anorexia

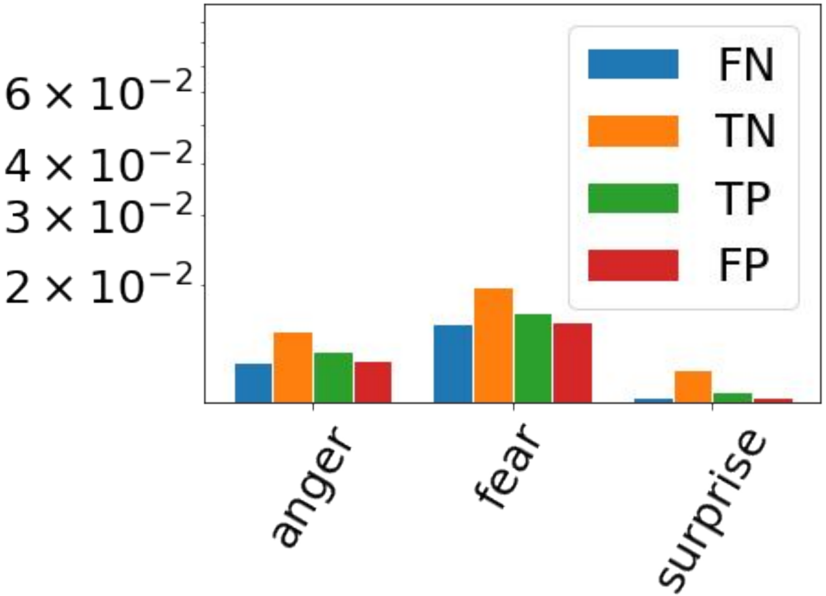
social, disgust, anxiety, feel, adverb, future, ingest, bio, health, pronoun, I, work, leisure, article, money,

PTSD

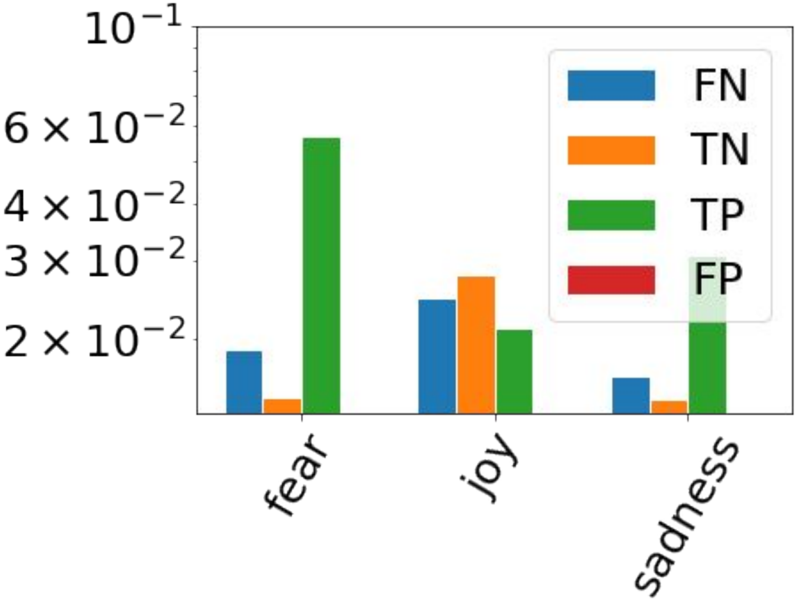
future, she/he, negative, anger, anxiety, health, sadness, fear, feel, anticipation, positive emotion,

Error analysis - emotions

Depression (reddit)

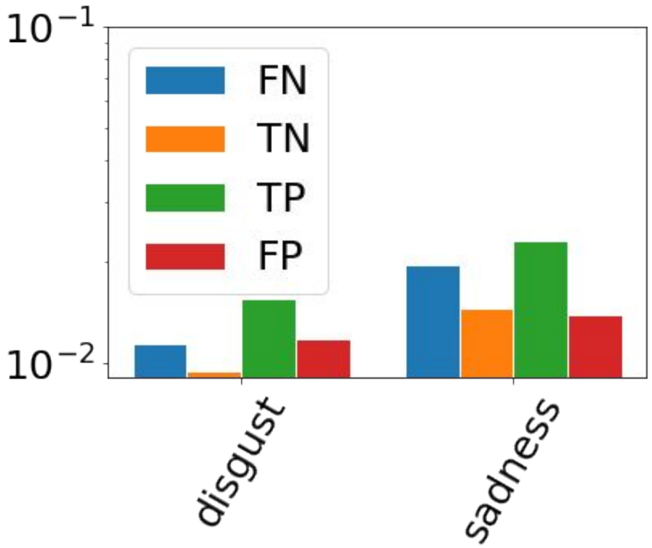


PTSD (Twitter)

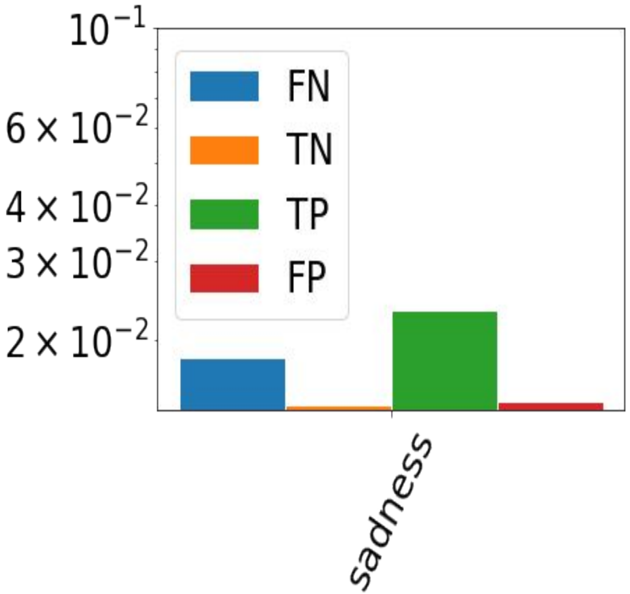


Error analysis - emotions

Anorexia (reddit)



Self-harm (reddit)



Emotion feature analysis - limitations

`Pretty much spamming it. Fe ling sucidal. I hope there was an earthquake today.`

Anger	feeling, earthquake	0.125
Anticipation	feeling, hope, pretty	0.25
Disgust	feeling	0.06
Fear	feeling, earthquake	0.125
Joy	feeling, hope, pretty	0.25
Negative	feeling, earthquake	0.125
Positive	feeling, hope, pretty	0.25
Sadness	feeling, earthquake	0.125
Surprise	feeling, hope, earthquake	0.25
Trust	feeling, hope, pretty	0.187

Attention activations - word level

LSTM_layer (LSTM)	(None, 256, 128)	117248	embedding_dropout[0][0]
attention (Dense)	(None, 256, 1)	129	LSTM_layer[0][0]
flatten (Flatten)	(None, 256)	0	attention[0][0]
activation (Activation)	(None, 256)	0	flatten[0][0]
repeat_vector (RepeatVector)	(None, 128, 256)	0	activation[0][0]
permute (Permute)	(None, 256, 128)	0	repeat_vector[0][0]
multiply (Multiply)	(None, 256, 128)	0	LSTM_layer[0][0] permute[0][0]
lambda (Lambda)	(None, 128)	0	multiply[0][0]
sent_repr_dropout (Dropout)	(None, 128)	0	lambda[0][0]

Attention activations - user level

```
                                                                  numerical_dense_layer_user[0][0]
                                                                  sparse_dense_layer_user[0][0]
__________________________________________________________________________________________________
LSTM_layer_user (LSTM)          (None, 50, 32)       25728       concatenate[0][0]
__________________________________________________________________________________________________
attention_user (Dense)          (None, 50, 1)        33          LSTM_layer_user[0][0]
__________________________________________________________________________________________________
flatten_1 (Flatten)             (None, 50)           0           attention_user[0][0]
__________________________________________________________________________________________________
activation_1 (Activation)       (None, 50)           0           flatten_1[0][0]
__________________________________________________________________________________________________
repeat_vector_1 (RepeatVector)  (None, 32, 50)       0           activation_1[0][0]
__________________________________________________________________________________________________
permute_1 (Permute)             (None, 50, 32)       0           repeat_vector_1[0][0]
__________________________________________________________________________________________________
multiply_1 (Multiply)           (None, 50, 32)       0           LSTM_layer_user[0][0]
                                                                  permute_1[0][0]
__________________________________________________________________________________________________
lambda_1 (Lambda)               (None, 32)           0           multiply_1[0][0]
__________________________________________________________________________________________________
user_repr_dropout (Dropout)     (None, 32)           0           lambda_1[0][0]
__________________________________________________________________________________________________
output_layer (Dense)            (None, 1)            33          user_repr_dropout[0][0]
==================================================================================================
Total params: 2,148,491
Trainable params: 2,148,491
Non-trainable params: 0
```

Attention activations: anorexic user

>>> the fact that they ve seen me naked

>>> it s hypocritical like modern feminism in general it s wrong when a guy does it but perfect when a woman does it s sad really feminism started out as such a good thing i have so much respect and for the original feminists the ones who fought for equality not domination and superiority

>>> i only feel hostile towards the fat people who are hostile towards me like the ones who say shit like real women have curves fuck those stupid skinny bitches and real men want a woman with meat on her only dogs go for bones if a person s going to insult my body i m going to give it back whenever an overweight person tells me go eat a big mac i will say go eat a salad with a light dressing on the side when one tells me i m too skinny for anyone to ever want me i will tell them they are too fat for anyone to ever want no one wants your bones poking them in bed nobody wants to be crushed under your fat folds in bed if a person wants to be unhealthy and die an early death that s their choice but if they think that gives them the right to talk shit about my healthy weight i will show them what it feels like

>>> sexual assault an anxiety disorder abuse and bullying

>>> male victims of domestic abuse and sexual assault

>>> stand i feel like the floor of the shower is gross because that s where all of the run off lands when you get in and the water starts off the sweat and stuff

>>> i d try to hatch it so i d have a chicken so it would lay more eggs so i d have a steady food source in the mean time i d look for a water source and a temporary food source to tide me over during the period of the chicken

Attention activations: depressed user

>>> wow thank you so much for this much needed response this made me smile so much i am doing my best and hopefully soon i ll find my happiness and it s true music is so strong in all aspects

>>> thank you i agree this game helped people in so many ways i guess it helped fill in a void

>>> i would never sell it it really is priceless d i don t know why anyone would but perhaps financial reasons it think that person is really lucky as well well true though i feel like they really wanted to release it as a merchandise but i heard it had something to do with the in music it s unfortunate though i think it would ve helped the company greatly since music is one of the selling points to the game

>>> wow really congrats d did you win it through a too

>>> the limited edition of the game is on ebay if you want the cd version of soundtrack it s a great deal

Attention activations: self-harming user

>>> i am on medication three types and xanax too to help with anxiety attacks i can tell that it helps but i m still struggling a lot i ll be talking again to my psychiatrist and psychologist in about weeks and see what they say i guess part of my anxiety is constantly seeing if i m alone in it and seeking reassurance it was an impulse to post

>>> any ideas i just cut yesterday but have to work tomorrow i cut my lower arms it s what helps the most but now i have to hide the evidence at work other than a long sleeve shirt is there anything else that might hide what i ve done thanks

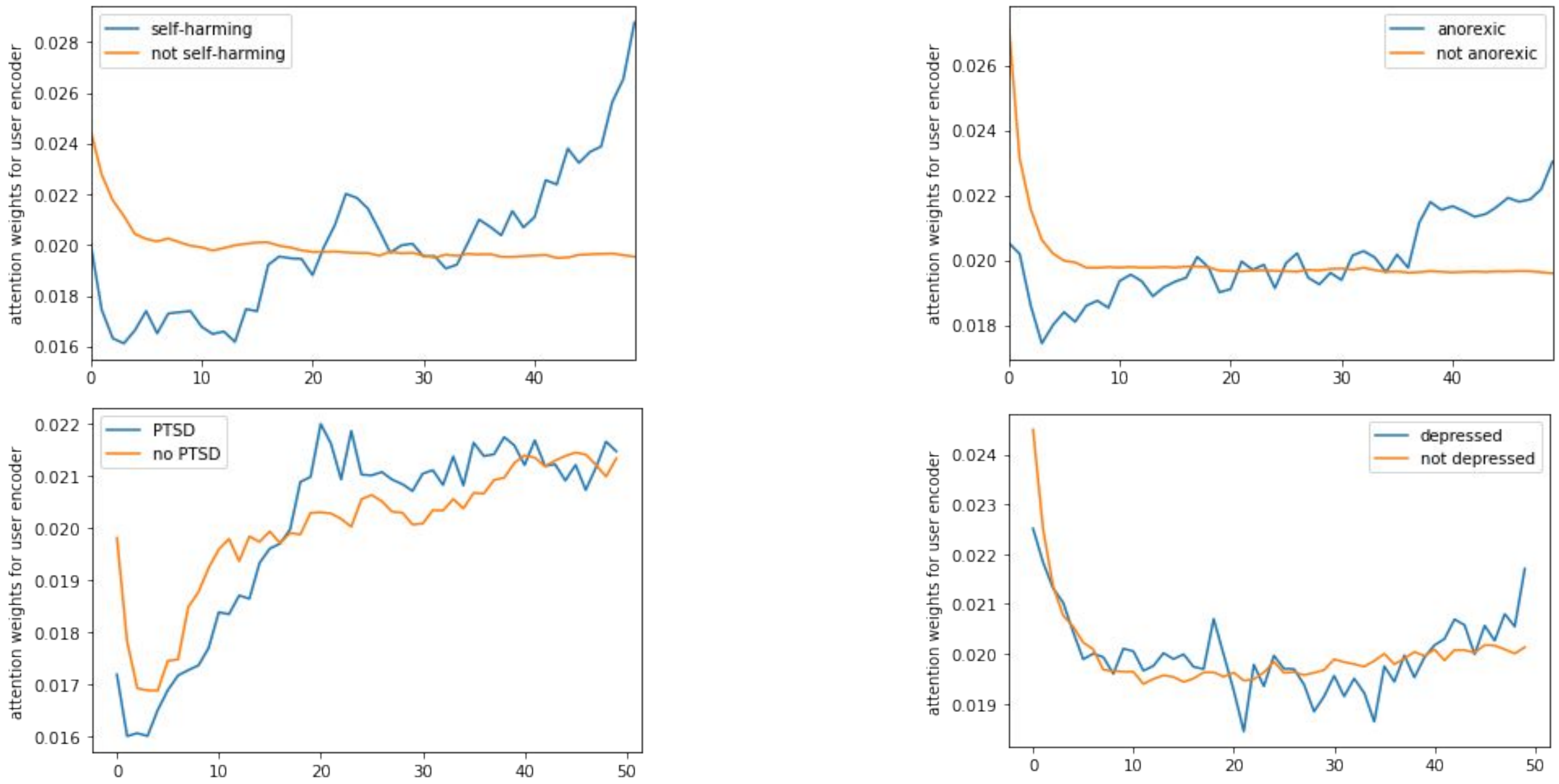
>>> i don t think you ever receive an e mail about it i would e mail them directly and explain your problem or you could wait a couple of days if you wanted i wouldn t but you might

>>> new jersey represent

>>> i m going to be that jerk that says i still don t like it well i mean its not that i don t like it s just not let s play if it was like just additional merch or certain new let s play only had it i d be fine with it

User-level attention - average distribution

Increasing importance over time

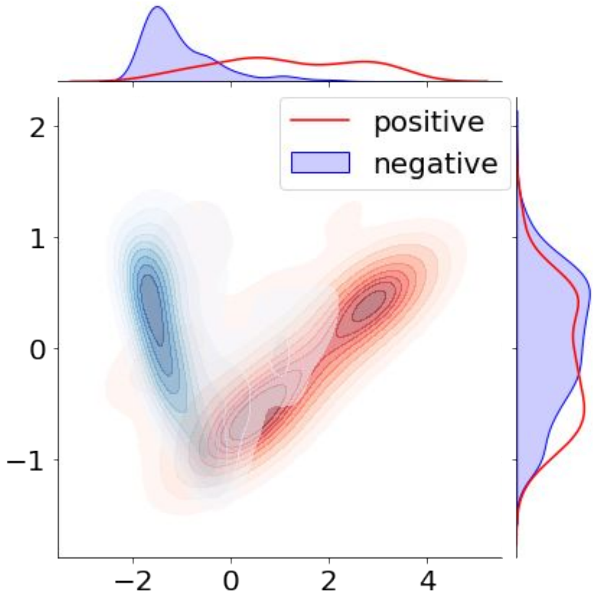


User embeddings

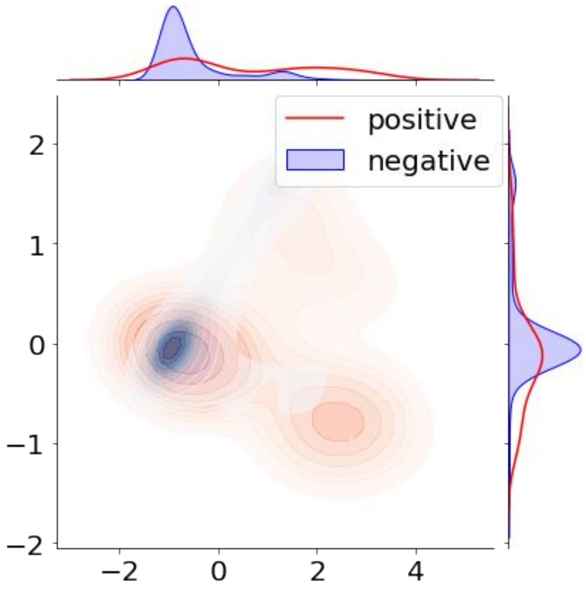
Hidden layer analysis

```
                                                                   numerical_dense_layer_user[0][0]
                                                                   sparse_dense_layer_user[0][0]
__________________________________________________________________________________________________
LSTM_layer_user (LSTM)          (None, 50, 32)       25728       concatenate[0][0]
__________________________________________________________________________________________________
attention_user (Dense)          (None, 50, 1)        33          LSTM_layer_user[0][0]
__________________________________________________________________________________________________
flatten_1 (Flatten)             (None, 50)           0           attention_user[0][0]
__________________________________________________________________________________________________
activation_1 (Activation)       (None, 50)           0           flatten_1[0][0]
__________________________________________________________________________________________________
repeat_vector_1 (RepeatVector)  (None, 32, 50)       0           activation_1[0][0]
__________________________________________________________________________________________________
permute_1 (Permute)             (None, 50, 32)       0           repeat_vector_1[0][0]
__________________________________________________________________________________________________
multiply_1 (Multiply)           (None, 50, 32)       0           LSTM_layer_user[0][0]
                                                                 permute_1[0][0]
__________________________________________________________________________________________________
lambda_1 (Lambda)               (None, 32)           0           multiply_1[0][0]
__________________________________________________________________________________________________
user_repr_dropout (Dropout)     (None, 32)           0           lambda_1[0][0]
__________________________________________________________________________________________________
output_layer (Dense)            (None, 1)            33          user_repr_dropout[0][0]
==================================================================================================
Total params: 2,148,491
Trainable params: 2,148,491
Non-trainable params: 0
```

User embeddings in 2D

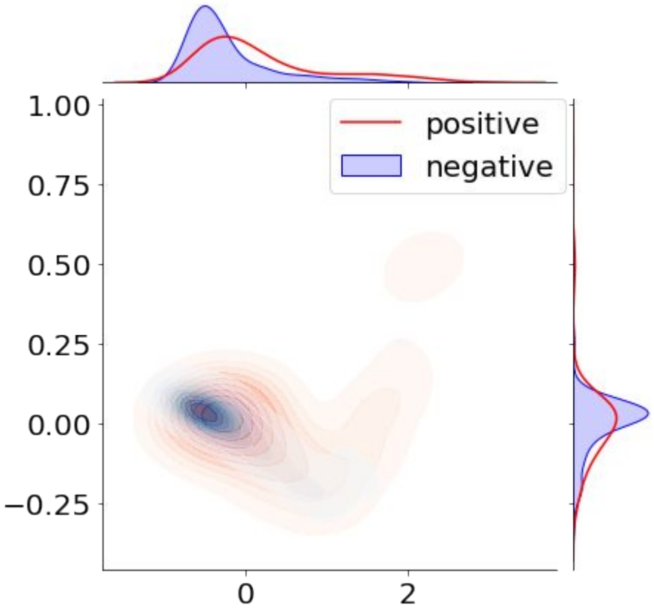


Anorexia (reddit)

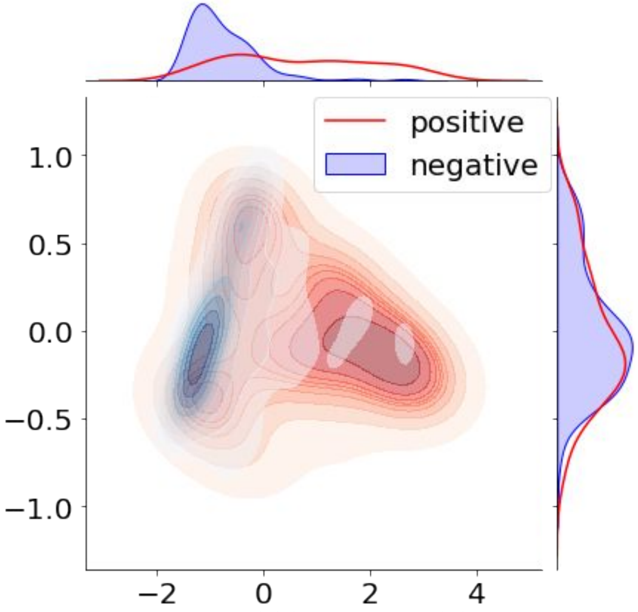


Depression (reddit)

User embeddings in 2D



PTSD (twitter)



Self-harm (reddit)

Clustering patterns of anorexia

(a) Cluster of users with anorexia **ANO1**.

(b) Cluster of users with anorexia **ANO2**.

(c) Cluster of control cases.

Feature analysis: Anorexia clusters

	ANO1	ANO2	Control
work**	1.22	1.47	2.31
money**	0.41	0.50	0.86
leisure**	1.12	1.15	1.88
pronoun***	17.41	16.20	11.54
I***	6.95	5.52	3.49
we*	0.33	0.46	0.51
friend**	0.22	0.25	0.15
family**	0.27	0.28	0.22
humans**	0.82	0.92	0.72

Table 4: Features about everyday activities and social relations, percentage of average usage per cluster.
*** Statistically significant difference across the three clusters
** Statistically significant difference between people suffering from anorexia and control users.
* Statistically significant difference between **ANO1** and others

Psycho-linguistic categories (LIWC)

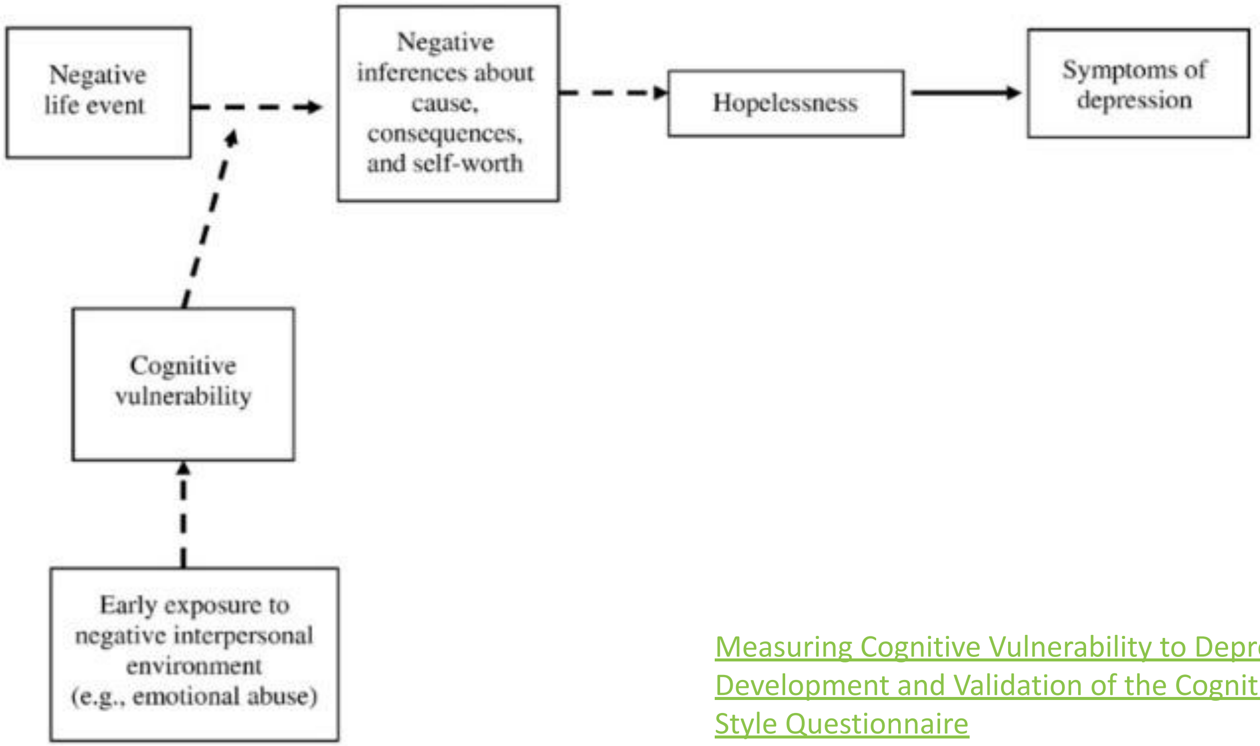
II. PSYCHOLOGICAL PROCESSES

Social Processes	talk, us, amigo
Friends	pal, buddy, coworker
Family	mom, brother, cousin
Humans	boy, woman, group
Affective Processes	happy, ugly, bitter
Positive Emotions	happy, pretty, good
Negative Emotions	hate, worthless, enemy
Anxiety	nervous, afraid, tense
Anger	hate, kill, pissed
Sadness	grief, cry, sad

Cognitive Processes	cause, know, ought
Insight	think, know, consider
Causation	because, effect, hence
Discrepancy	should, would, could
Tentative	maybe, perhaps, conjetura
Certainty	always, never
Inhibition	block, constrain
Inclusive	with, and, include
Exclusive	but, except, without

Cognitive styles

The hopelessness theory of depression



[Measuring Cognitive Vulnerability to Depression: Development and Validation of the Cognitive Style Questionnaire]

Feature analysis: Anorexia clusters

	ANO1	ANO2	Control
cogmech* **	16.43	15.88	14.58
feel **	0.86	0.86	0.43
certain **	1.55	1.69	1.04
tentative **	3.09	3.01	3.13
causation*	1.71	1.85	1.87

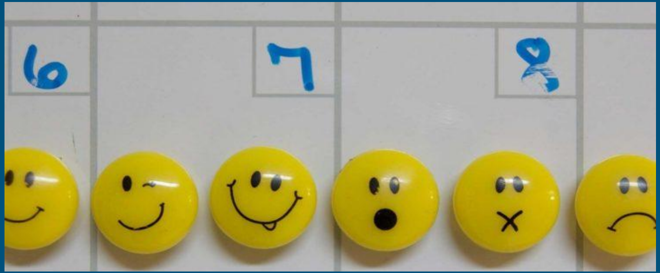
Table 5: Features about cognitive styles (cognitive processes and perceptual processes), percentage of average usage per cluster.

*** Statistically significant difference across the three clusters

** Statistically significant difference between people suffering from anorexia and control users.

* Statistically significant difference between **ANO1** and others

Emotions over time



Not only the static expression of certain emotions or discussion of topics is relevant, but their **evolution over time**

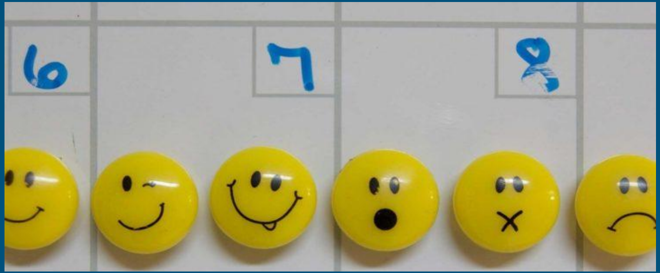
Track evolution of emotion expression over time

Track evolution of usage of different psycho-linguistic categories over time (LIWC)

Analyze their correlations

=> Understand how emotions relate to different psycho-linguistic categories (e.g. causation, society, self etc) for users suffering from a mental disorders

Emotions over time



Method:

Measure emotion usage in texts posted per day - separately for positive vs negative and users + average across users

Measure psycho-linguistic categories usage per day for each user, ...

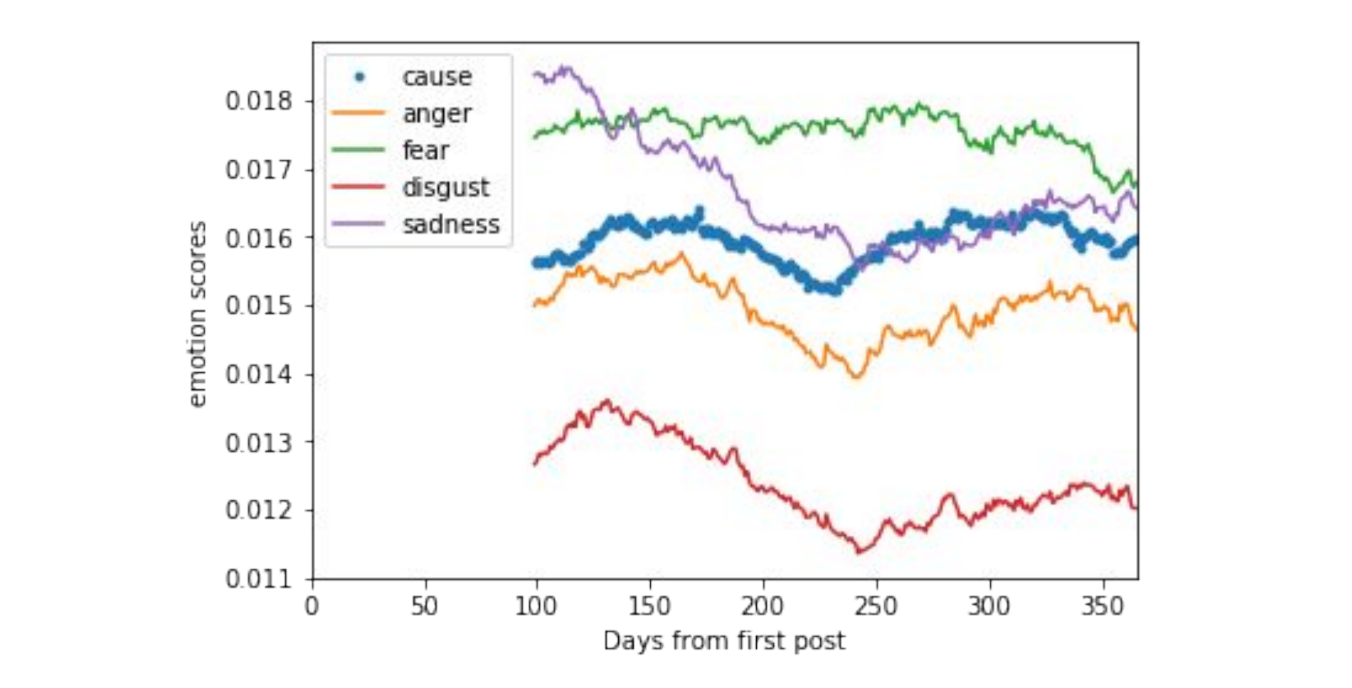
Rolling average of 100 days

Pearson correlations between obtained time series for every (emotion, psycho-linguistic category) pair

Compare correlations between positive users and negative users

Select pairs with significantly different correlations between the two groups (z-test)

Emotions over time



Depressed users expressing causation & negative emotions over time

Emotions over time: findings

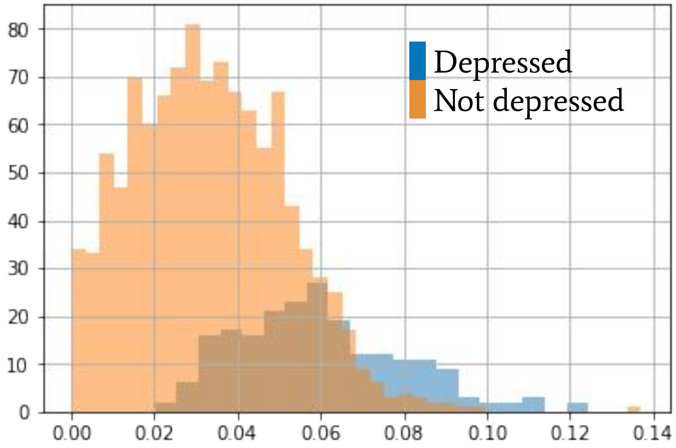
Causation and emotions

	Anger		Disgust		Fear		Sadness		Trust		Anticip.		Joy	
pos(P)/neg(N)	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**
Depression	0.39	-0.36	0.33	-0.21	0.46	-0.43	-	-	-0.06	-0.31	-	-	-0.23	-0.03
Anorexia	0.48	0.07	0.40	0.02	0.39	0.04	0.45	-0.38	0.41	0.16	-0.13	-0.23	-0.08	-0.55
Self-harm	0.25	0.12	-	-	0.15	0.27	-0.15	0.03	-	-	0.23	-0.17	0.26	-0.19

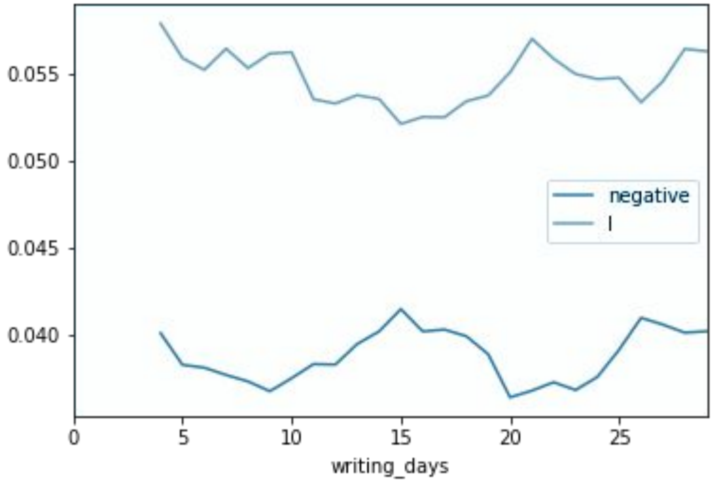
Table 7: Correlation between "causation" and emotions in the three mental disorders for positive users (diagnosed with a mental disorders) and negative ones (healthy). Only correlations which are significantly different between the positive and negative classes are shown.

Feature analysis (over time) - “I”

Depression



The use of “I” in depressed vs non-depressed users



Use of “I” vs negative emotion in depressed users

Emotions over time: findings

The self and emotions

	Anger		Disgust		Fear		Sadness		Trust		Anticip.		Joy	
pos(P)/neg(N)	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**
Depression	-	-	-	-	-0.11	-0.29	0.25	-0.04	0.15	-0.15	0.58	-0.62	0.50	-0.06
Anorexia	0.12	-0.16	0.08	-0.22	0.30	-0.16	0.24	0.06	0.27	-0.25	0.53	0.24	0.72	0.55
Self-harm	0.42	0.01	0.34	0.13	0.21	-0.28	0.34	-0.06	-	-	-0.16	0.31	-0.05	0.40

Table 8: Correlation between the use of "I" and emotions in the three mental disorder for positive users (diagnosed with a mental disorder) and negative ones (healthy). Only correlations which are significantly different between the positive and negative classes are shown.

Other tasks

- Detecting the **severity** of depression / suicide risk level
- Detecting specific symptoms (lack of sleep, loss of appetite, lack of energy...)
- Detecting **causes** of depression - helps with prevention, and with targeted management
- Detecting depression from video therapy sessions (based on video/audio signals)
- Analyze different disorders jointly (co-morbidities); transfer learning
- Profiling users suffering from a disorder: age, behavioral patterns, social media activity patterns (nocturnal, seasonal)
- Conversational data: therapy sessions, therapist chatbot (https://woebothealth.com/)
- Multimodal depression detection
- Social media: depression and **aggression**

In practice: eRisk 2021

Best results in overall level of depression prediction (some metrics) at Task 3:
http://ceur-ws.org/Vol-2936/paper-75.pdf

Transfer learning

Clinical evidence of comorbidity within mental disorders. (Exploring Comorbidity Within Mental Disorders Among a Danish National Population)

- Improve performance on tasks with less data (depression → other disorders)
- Understand connection/compatibility between disorders and expression media (genre/platform)

Cross-task - transfer knowledge between labels for different disorders

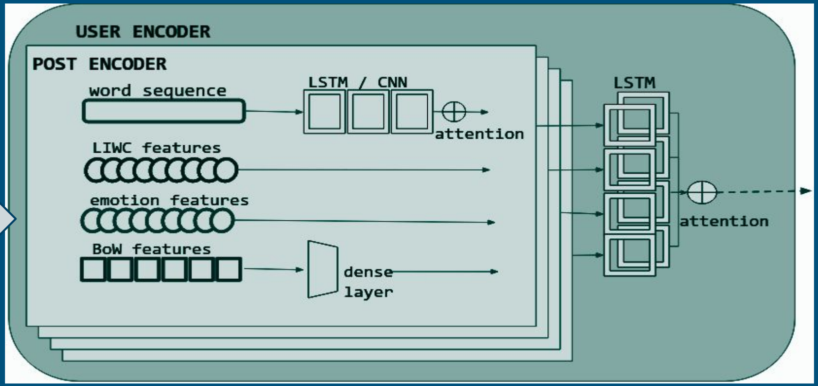
Cross-genre - transfer knowledge between different data platforms (reddit/Twitter)

-

Transfer learning

Strategy 0. No pre-training

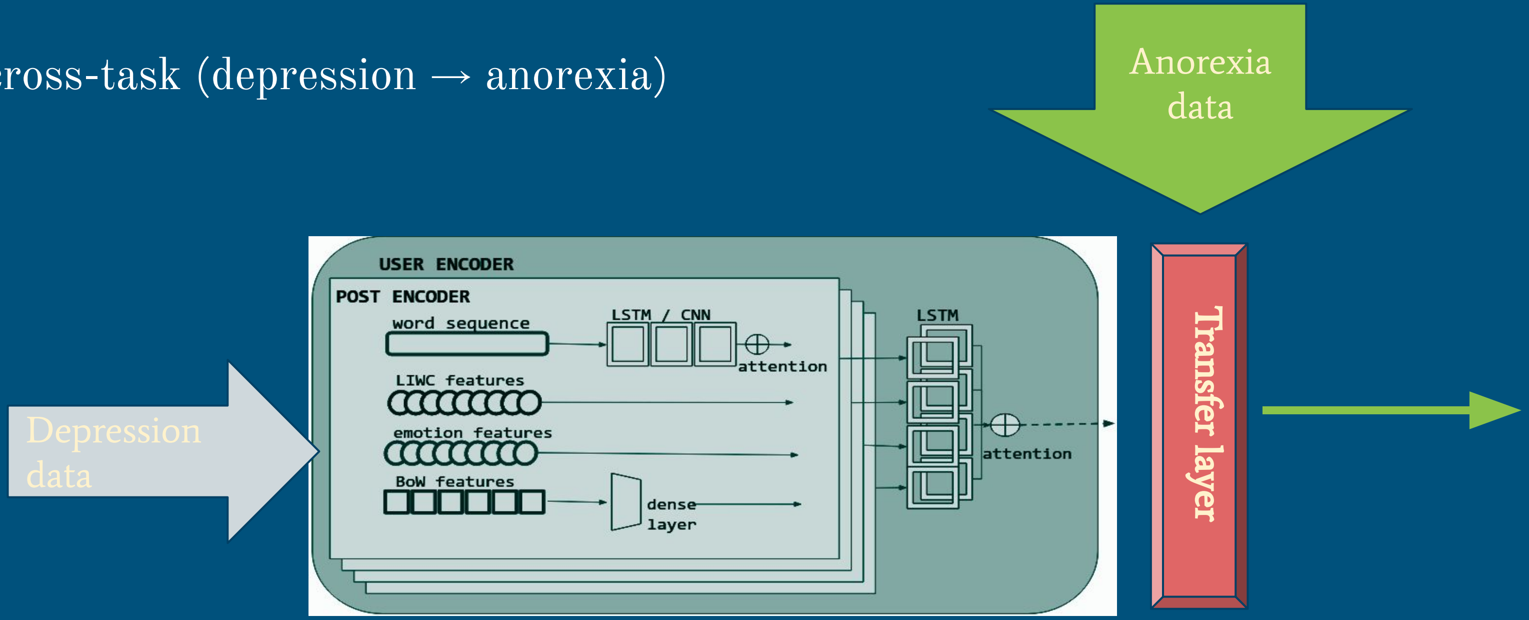
Depression data



Transfer learning

Strategy 1. Transfer layer

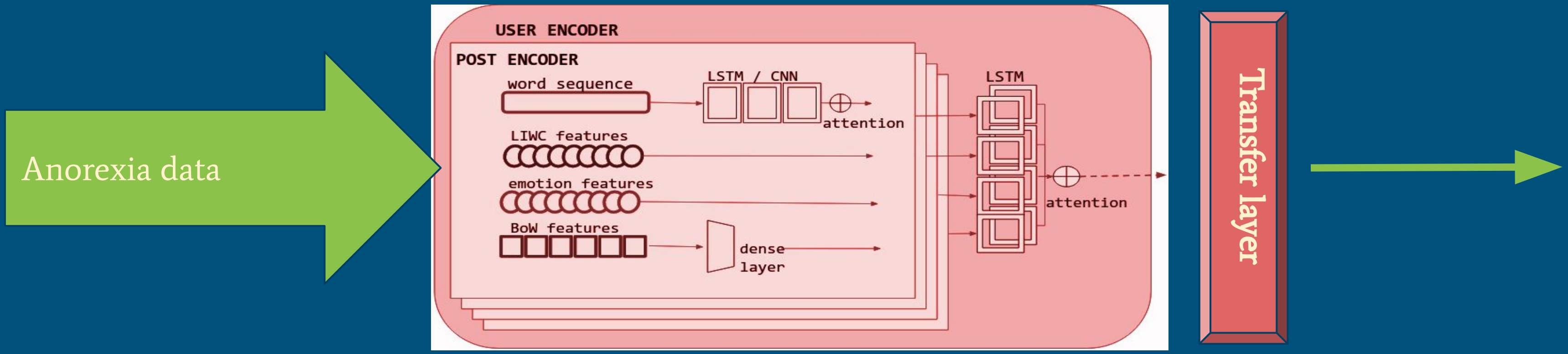
Example: cross-task (depression → anorexia)



Transfer learning

Strategy 2. Fine-tuning

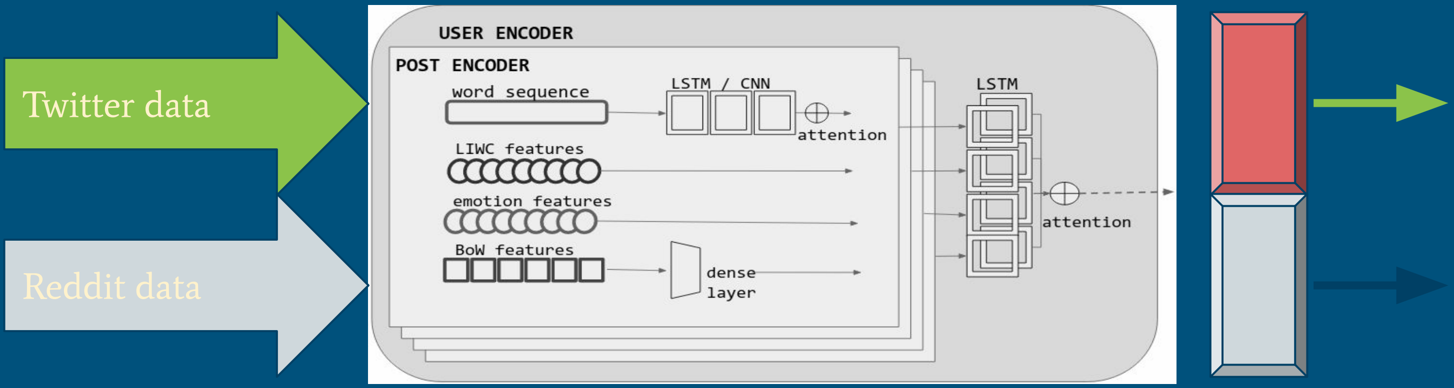
Example: cross-task (depression → anorexia)



Transfer learning

Strategy 3. Multi-task learning

Example: cross-genre (reddit / Twitter)



Transfer learning experiments: Results

	CROSS-TASK						CROSS-GENRE			
Source	eRisk depression				CLPsych depression		eRisk depression			
Target	eRisk Anorexia		eRisk Self-harm		CLPsych PTSD		(Shen et al.) depression		CLPsych depression	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
Strategy 0	.17	.62	.13	.69	.31	.60	.69	.59	.38	.57
Strategy 1	.64	.90	.54	.87	.43	.73	.65	.74	.61	.72
Strategy 2	.63	.93	.67	.87	.58	.78	.86	.94	.60	.74
Baseline BiLSTM	.62	.93	.62	.84	.55	.78	.75	.83	.56	.72

Source	All depression					
Target	eRisk		(Shen et al.)		CLPsych	
	F1	AUC	F1	AUC	F1	AUC
Strategy 3	.39	.81	.74	.83	.56	.82
Single-task	.40	.83	.75	.83	.56	.72