
Predicting trigger warning labels
Predicting trigger warning labels
Predicting trigger warning labels
Predicting trigger warning labels
Predicting trigger warning labels
Predicting trigger warning labels
Jamie Davis

PROBLEM
PROBLEM
PROBLEM
PROBLEM
PROBLEM

- Social media provide platforms for seeking peer-support with mental health.
- Platforms are largely unmoderated, potentially exposing users to dangerous content.
- Moderation is key to promoting the growth of safe support networks online.
- Unassisted manual moderation is infeasible.

AIM

Develop a model that **suggests trigger warnings** to assign to **posts** to facilitate content moderation and protect users.

- Problem can be framed as a text classification task.
- Use SpaCy v3 to train and deploy a multiclass text classification
- Model will be able to predict trigger warnings for free-text posts.

METHOD
METHOD
METHOD
METHOD
METHOD

Data

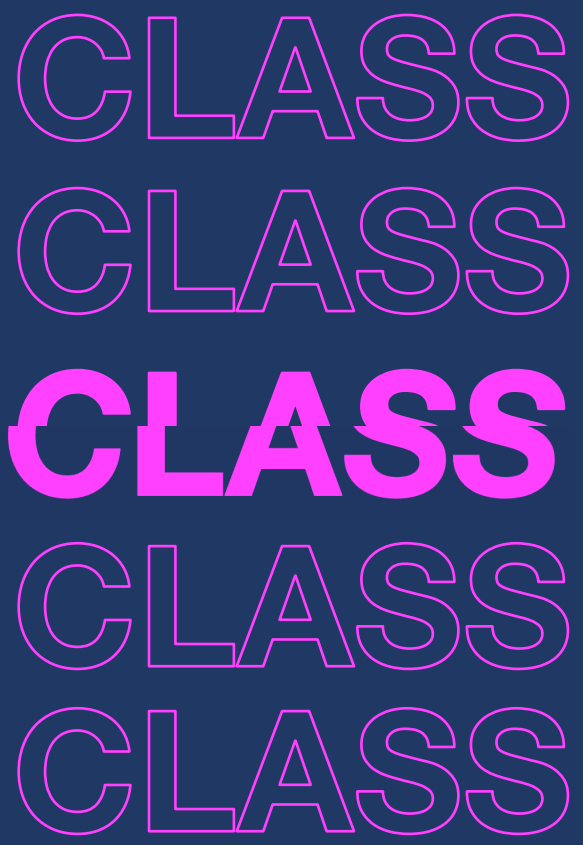
- Leveraged PushShift API to scrape 25,000 documents per subreddit.
- Each subreddit focused on a different mental health condition.
- After cleaning just over 142,000 documents remained.
- Split data:

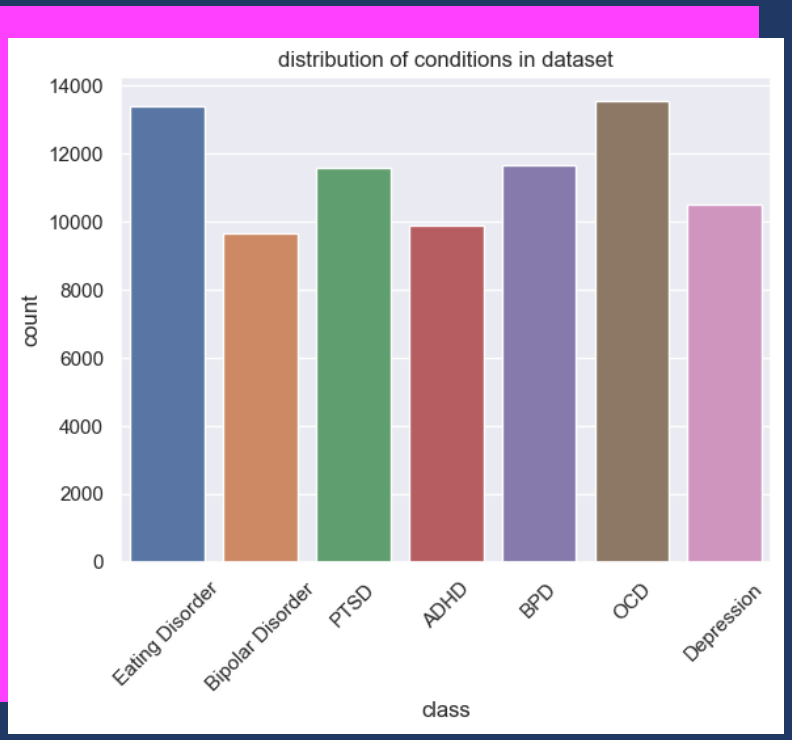
Train	valid	test
0.56	0.19	0.25

METHOD
METHOD
METHOD
METHOD
METHOD

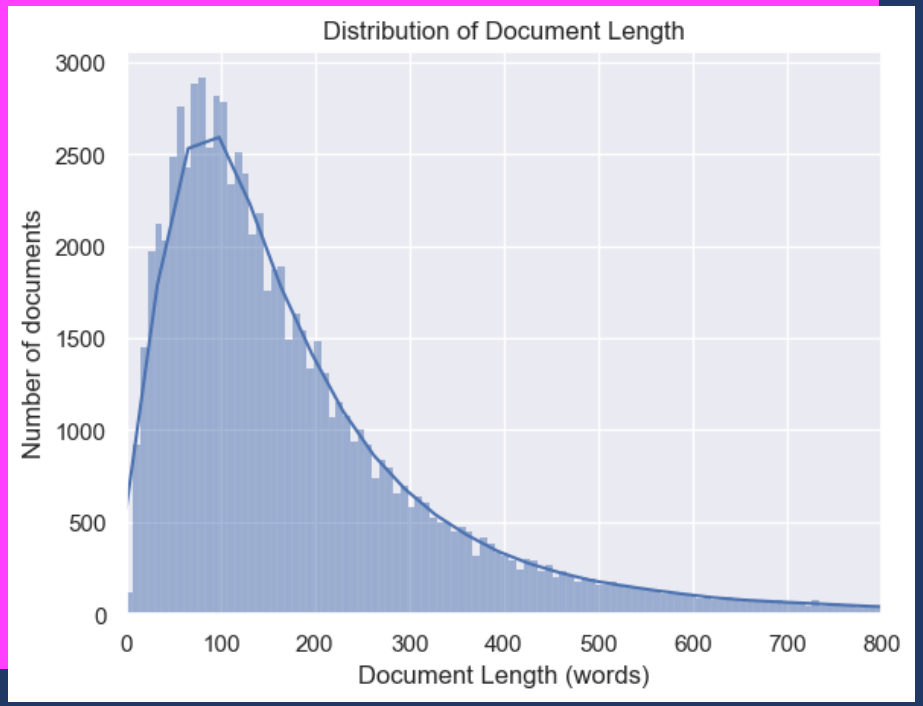
EXPLORATORY DATA ANALYSIS

- Depression
- Post Traumatic Stress Disorder (PTSD)
- Obsessive Compulsive Disorder (OCD)
- Borderline Personality Disorder (BPD)
- Attention Deficit Hyperactivity Disorder (ADHD)
- Bipolar Disorder
- Eating Disorder





CLASS
CLASS
CLASS
CLASS
CLASS



LENGTH
LENGTH
LENGTH
LENGTH
LENGTH

ADHD		BPD	
adhd	420.9	bpd	407.9
day	267.6	friend	337.8
work	244.6	people	330.6
medication	220.6	relationship	274.5
med	210.6	feeling	270.1

OCD		Depression	
ocd	708.1	life	404.4
thought	574.8	friend	288.9
intrusive	291.3	people	287.4
help	262.5	day	275.3
people	252.6	year	271.3

ED		PTSD	
weight	421.5	ptsd	415.2
eat	392.9	year	301.8
eating	355.4	trauma	279.6
food	344.5	people	264.6
ed	326.5	help	254.3

Bipolar	
bipolar	295.9
episode	230.6
med	221.8
day	221.6
manic	206.1

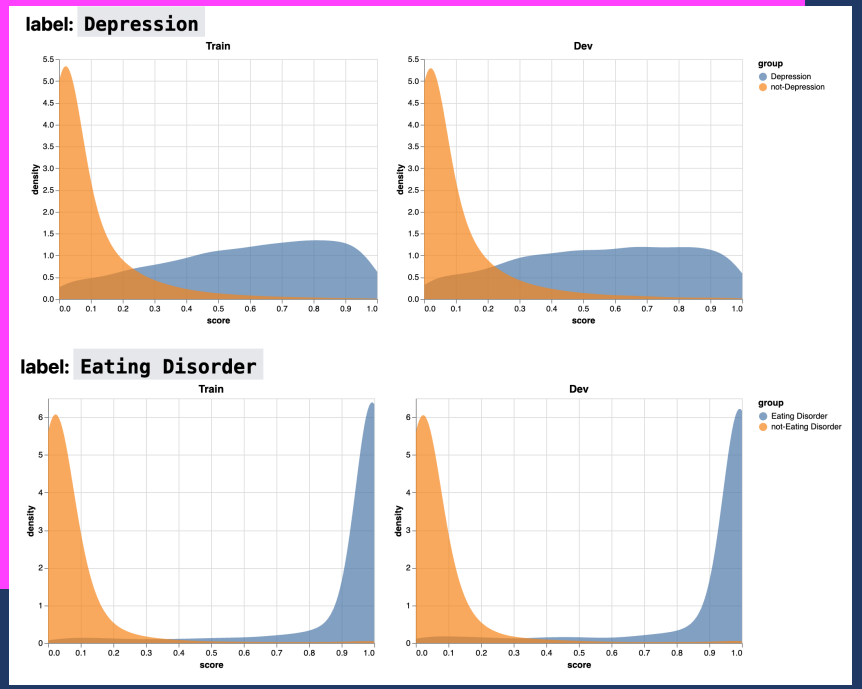
- Most documents seem to be around 70-100 words, sufficient context no need for pretrained vectors.
- ED and OCD have the most observations, bipolar and ADHD have the least – resulting model may be better at predicting the former?
- ADHD and Depression top words seem to be quite generic compared to others - will model reflect this?

- Convolutional neural network (CNN).
- Less accurate than ensemble or transformer-based model.
- Runs faster as it is less computational expensive

RESULTS
RESULTS
RESULTS
RESULTS
RESULTS
RESULTS

```
================================== Results ==================================

TOK                 100.00
TEXTCAT (macro F)   80.79
SPEED               582771


============================ Textcat F (per label) ============================

                        P       R       F
Depression          71.99   65.62   68.66
PTSD                85.63   76.79   80.97
OCD                 96.40   79.52   87.15
Eating Disorder     91.72   89.57   90.63
Bipolar Disorder    81.78   69.35   75.05
BPD                 84.61   69.92   76.57
ADHD                87.64   85.37   86.49


======================== Textcat ROC AUC (per label) ========================

                    ROC AUC
Depression             0.94
PTSD                   0.96
OCD                    0.97
Eating Disorder        0.99
Bipolar Disorder       0.95
BPD                    0.94
ADHD                   0.98
```

RESULTS
RESULTS
RESULTS
RESULTS
RESULTS



RESULTS
RESULTS
RESULTS
RESULTS
RESULTS
RESULTS

- **Macro F1-score (96%)** suggests model is very good at distinguishing classes
- **High precision (96%) for OCD** – very few false positives.
- **High recall and F1 for ED** – misses very few true positives for ED and overall is most distinguishable condition.
- Model struggled most with depression.

NEXT

NEXT

NEXT

NEXT

NEXT

- Adding a NER component to extract more detail.
- Test model on data from different sources.
- Add more categories to model.