Capstone Project: NLP, Deep Learning, & Reddit

Eileen Hartnett
General Assembly

01

Introduction

Why look at Reddit text to study mental illness?

128

r/depression · Posted by u/Anders-94 8 hours ago 2

I've never been so depressed in my life.

I'm not making this because I need sympathy, I'm making this cause I think I need help.

3

r/AnorexiaNervosa · Posted by u/moonring_ 1 day ago

Really struggling with recovery and didn't know where else to turn

Recovery Related

Vote

r/schizophrenia · Posted by u/mowyourass 1 hour ago

I am experiencing almost all of these on the list, but I am scared of hospitals and my family's judgement. How do you know that this is the right decision?

Need Support

SCIENTIFIC REPORTS
natureresearch

Check for updates

OPEN

A deep learning model for detecting mental illness from user content on social media

Jina Kim[1,2], Jieon Lee[1], Eunil Park[1,3]✉ & Jinyoung Han[3]✉

Users of social media often share their feelings or emotional states through their posts. In this study, we developed a deep learning model to identify a user's mental state based on his/her posting information. To this end, we collected posts from mental health communities in *Reddit*. By analyzing and learning posting information written by users, our proposed model could accurately identify whether a user's post belongs to a specific mental disorder, including depression, anxiety, bipolar, borderline personality disorder, schizophrenia, and autism. We believe our model can help identify potential sufferers with mental illness based on their posts. This study further discusses the implication of our proposed model, which can serve as a supplementary tool for monitoring mental health states of individuals who frequently use social media.

TABLE OF CONTENTS

02

PREPROCESSING

Preprocessing steps

Web scraping Reddit posts

Data cleaning: Tokenization, Stop Words, Porter Stemmer

NLTK
This was used for the XGBoost models

Word2Vec
This was used for the CNN models

Pre-Processing in depth

Scrape Reddit Posts

r/mentalhealth
r/depression
r/Anxiety
/bipolar
r/BPD
r/schizophrenia
r/autism
r/AnorexiaNervosa
r/Bulimia

Clean Text

-Remove punctuation
-Stop words
-Numbers
-non-English characters
- Tokenize
-Porter-Stem

EDA

- 84,879 posts
- Roughly even amounts across subreddits
- “PreCovid” shutdown: 41,955
- “Post-Covid” shutdown: 42,924

NLTK

Word2Vec

03

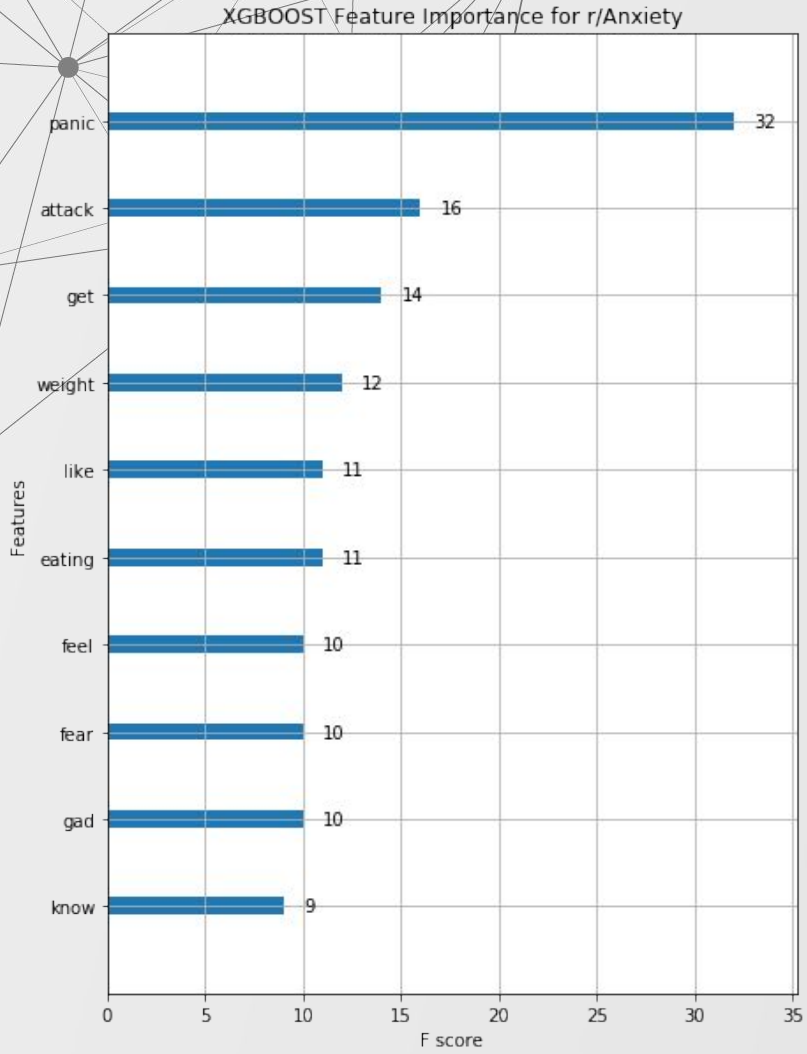
Analysis 1: XGBoost

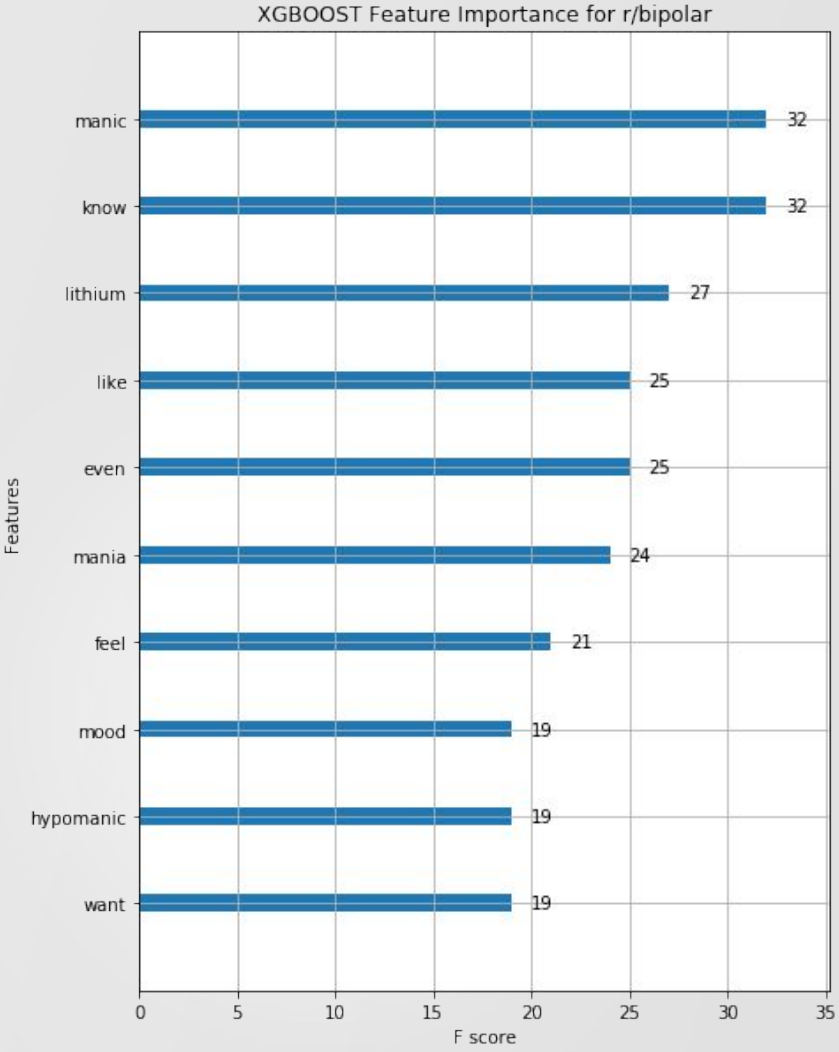
Kim et al. XGBoost Results

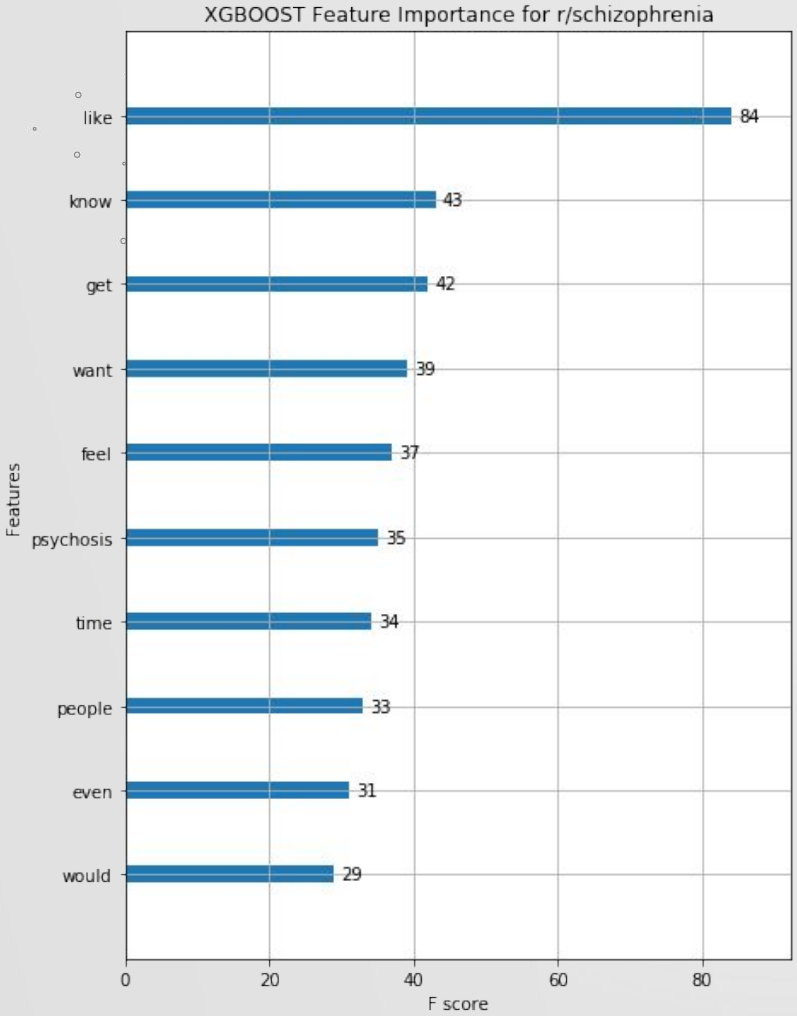
Channel	Class	F1-Score	Accuracy
r/depression	Depression	58.02	71.69
	Non-depression	78.65	
r/Anxiety	Anxiety	55.92	70.41
	Non-anxiety	77.73	
r/bipolar	Bipolar	53.59	85.53
	Non-bipolar	91.43	
r/BPD	BPD	46.43	85.14
	Non-BPD	91.37	
r/schizophrenia	Schizophrenia	40.97	86.72
	Non-schizophrenia	92.52	
r/autism	Autism	38.31	94.91
	Non-autism	97.35	

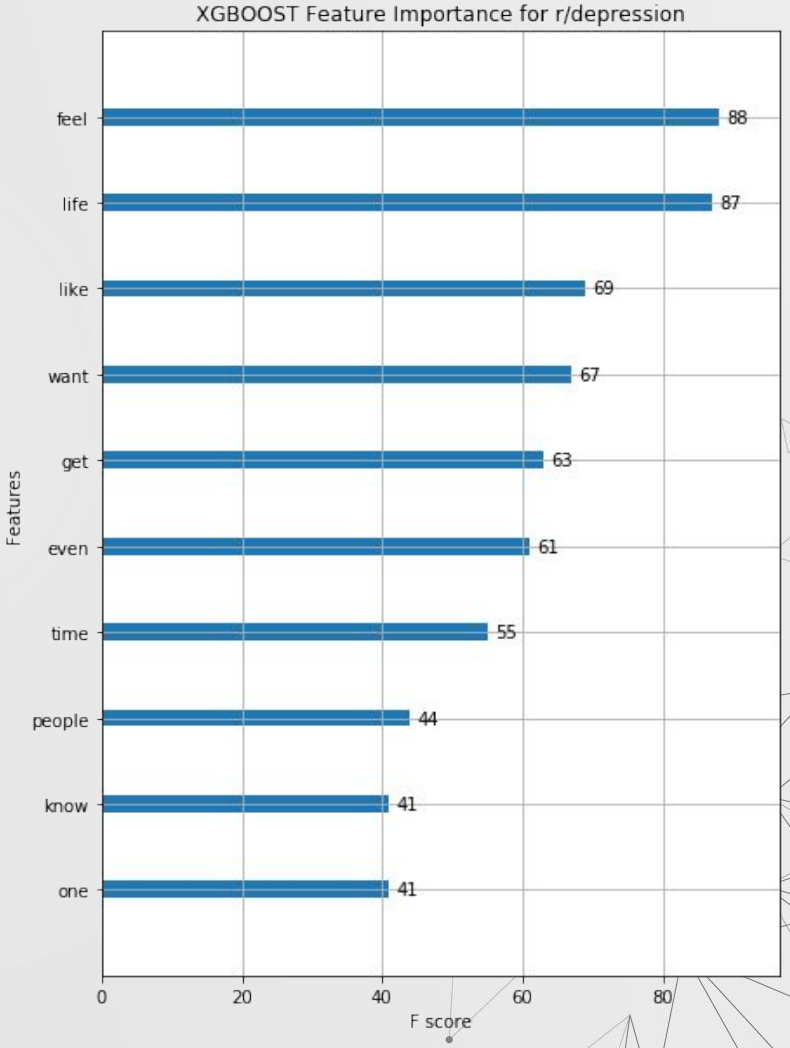
My XGBoost Results

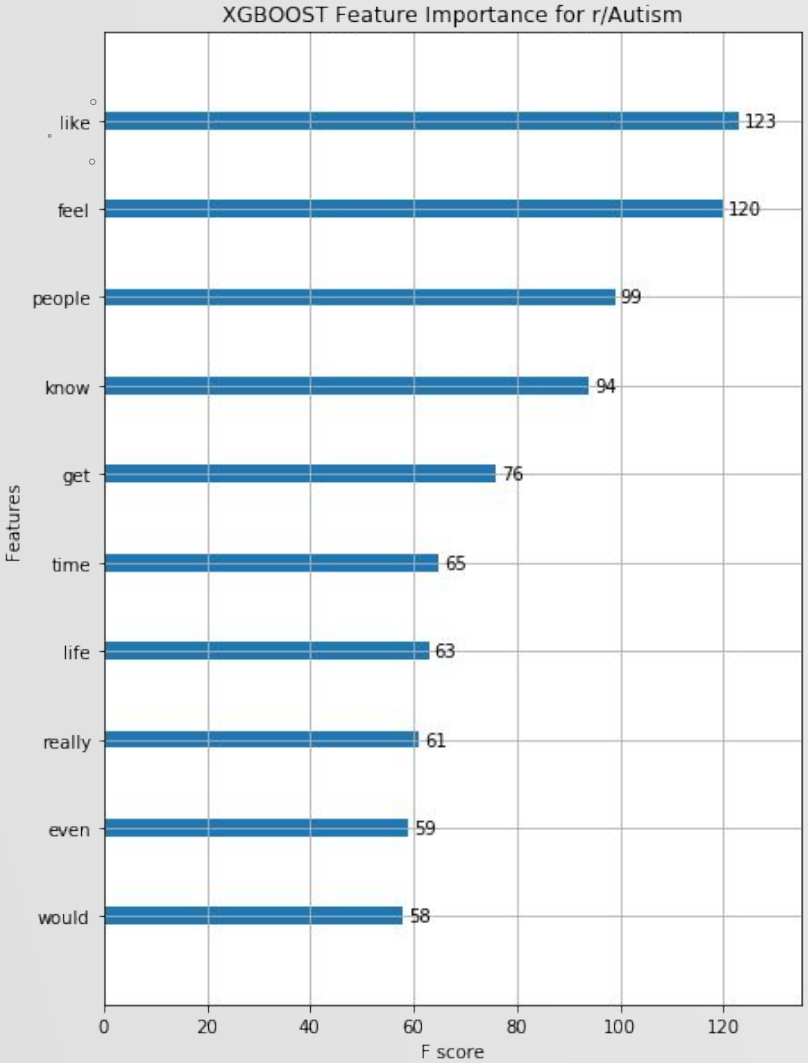
Channel	Class	F1-Score	Accuracy
r/depression	Depression	23	74.24
	Non-depression	93	
r/Anxiety	Anxiety	17	65.52
	Non-anxiety	92	
r/bipolar	Bipolar	34	72.25
	Non-bipolar	82	
r/BPD	BPD	28	71.77
	Non-BPD	82	
r/schizophrenia	Schizophrenia	37	77.83
	Non-schizophrenia	87	
r/autism	Autism	48	83.15
	Non-autism	90	

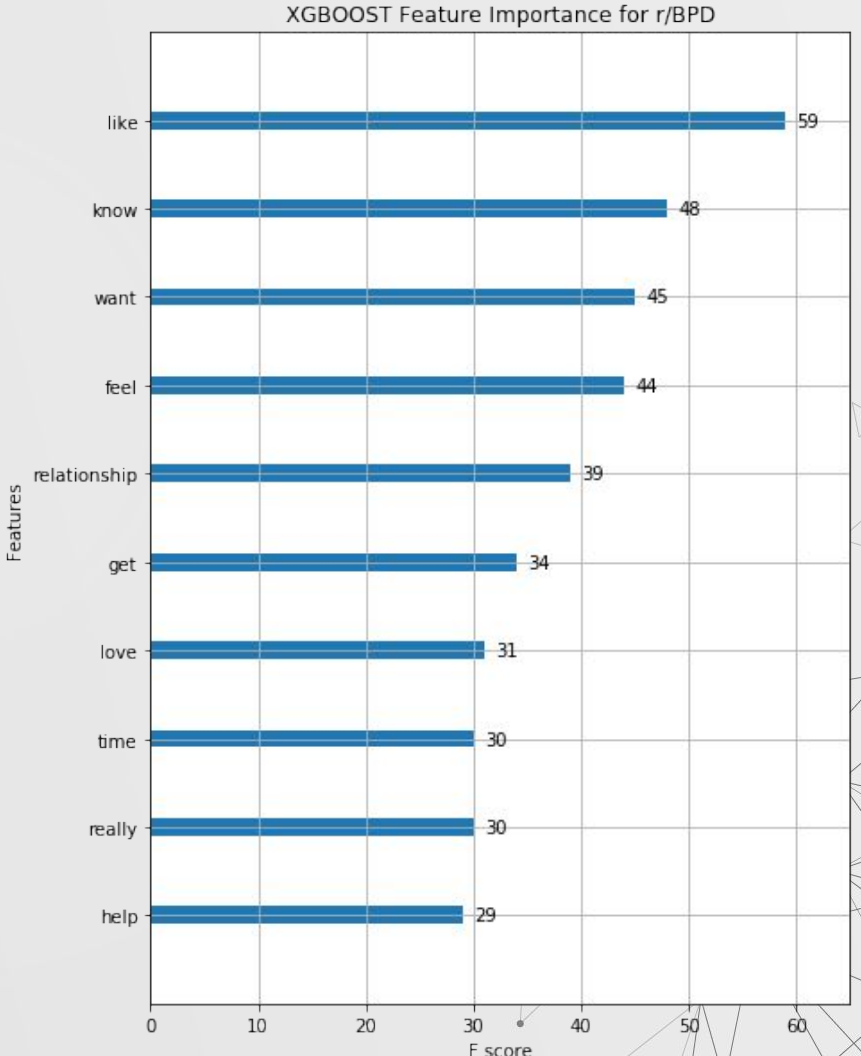

XGBOOST Feature Importance for r/Anxiety
panic
attack
get
weight
like
eating
feel
fear
gad
know
32
16
14
12
11
11
10
10
10
9
Features
0
5
10
15
20
25
30
35
F score

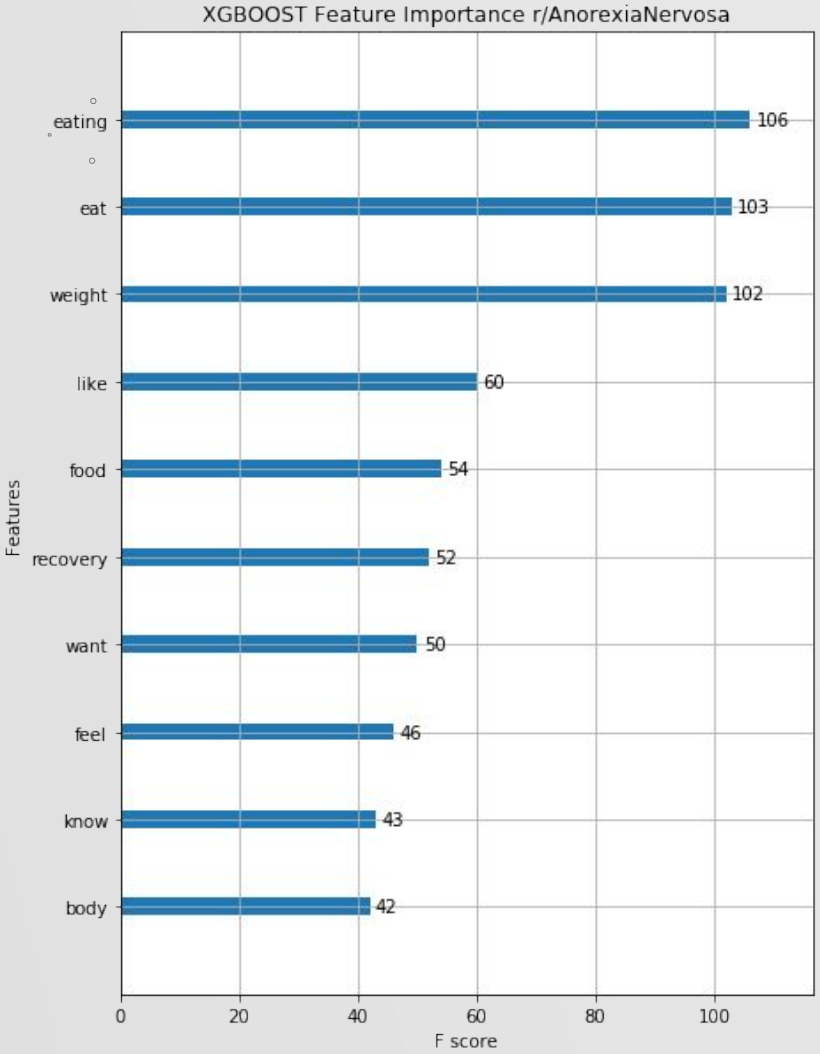

XGBOOST Feature Importance for r/bipolar
manic
know
lithium
like
even
mania
feel
mood
hypomanic
want
32
32
27
25
25
24
21
19
19
19
Features
0
5
10
15
20
25
30
35
F score

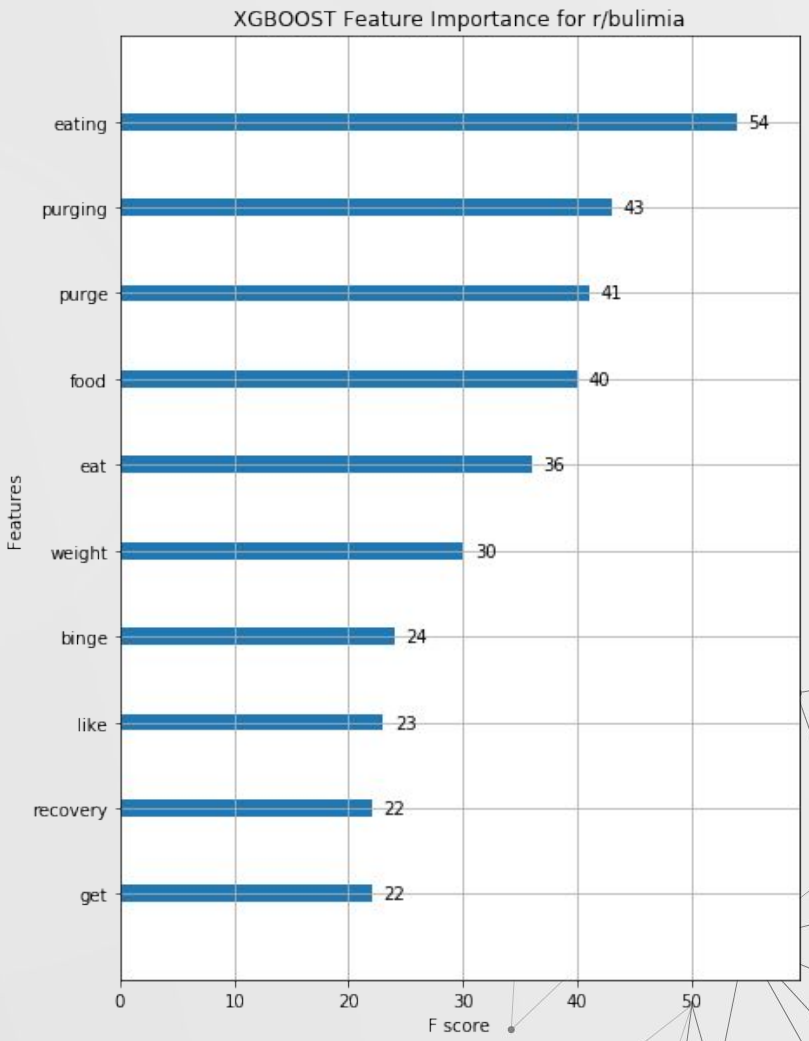

XGBOOST Feature Importance for r/schizophrenia
like
know
get
want
feel
psychosis
time
people
even
would
84
43
42
39
37
35
34
33
31
29
Features
0
20
40
60
80
F score


XGBOOST Feature Importance for r/depression
feel
life
like
want
get
even
time
people
know
one
88
87
69
67
63
61
55
44
41
41
Features
0
20
40
60
80
F score


XGBOOST Feature Importance for r/Autism
like
feel
people
know
get
time
life
really
even
would
123
120
99
94
76
65
63
61
59
58
Features
0
20
40
60
80
100
120
F score


XGBOOST Feature Importance for r/BPD
like
know
want
feel
relationship
get
love
time
really
help
59
48
45
44
39
34
31
30
30
29
Features
0
10
20
30
40
50
60
F score


XGBOOST Feature Importance r/AnorexiaNervosa
eating
eat
weight
like
food
recovery
want
feel
know
body
106
103
102
60
54
52
50
46
43
42
Features
0
20
40
60
80
100
F score


XGBOOST Feature Importance for r/bulimia
eating
purging
purge
food
eat
weight
binge
like
recovery
get
54
43
41
40
36
30
24
23
22
22
Features
0
10
20
30
40
50
F score

04

Analysis 2: Word2Vec & CNN

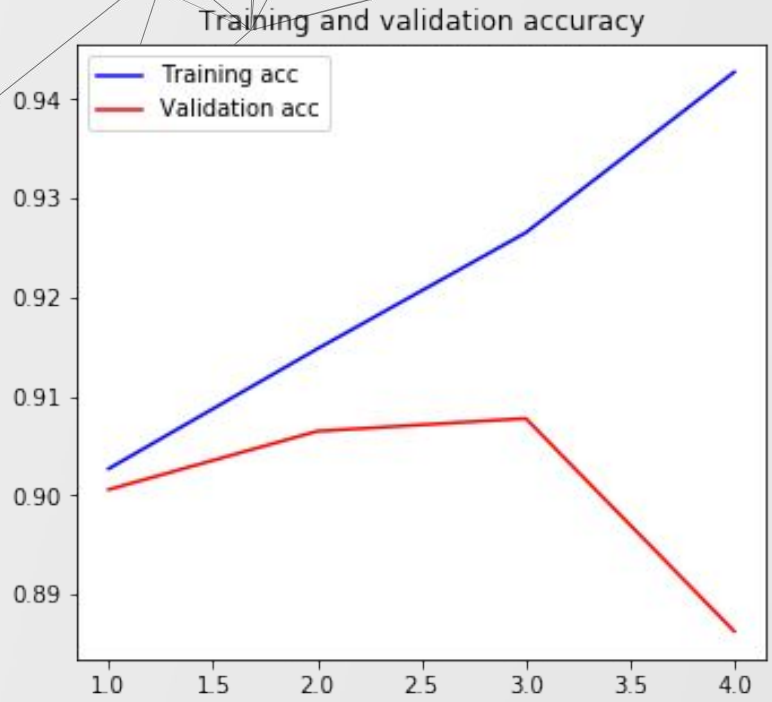
Word2Vec

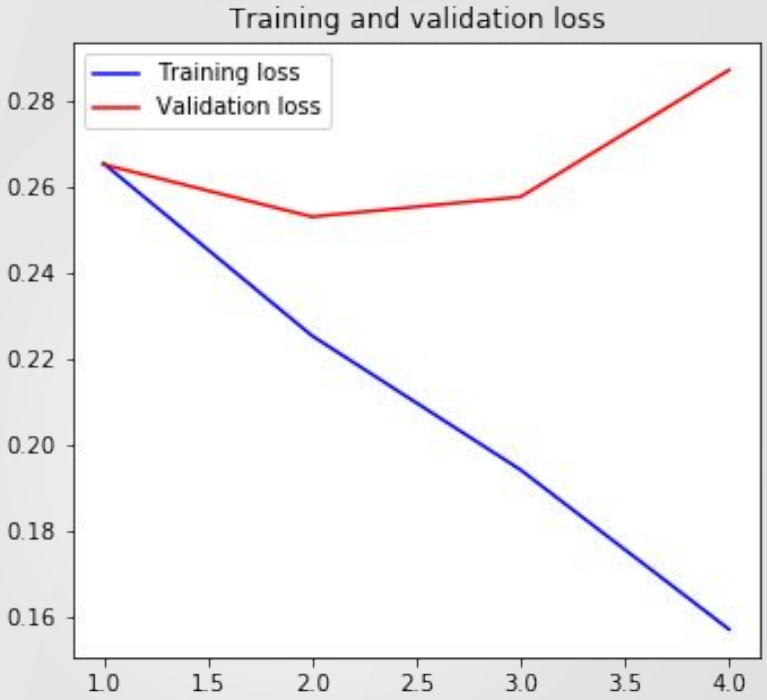
Gif credit: https://moderndata.plotly.com/understanding-word-embedding-arithmetic/

CNN Accuracy Results

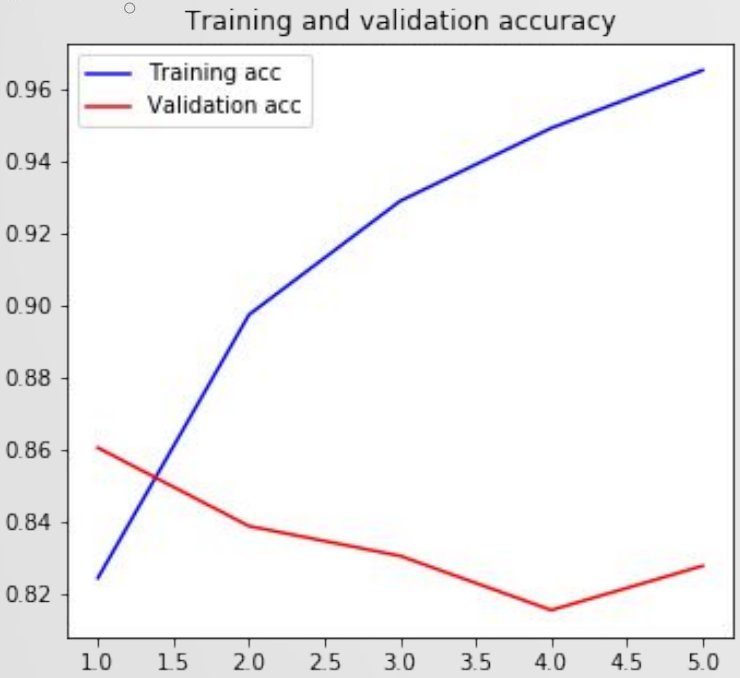
	r/Anxiety	r/Anorexia	r/Schizophrenia
Test Accuracy	82.76%	83.09%	79.14%
Train Accuracy	98.14%	98.15%	97.53%

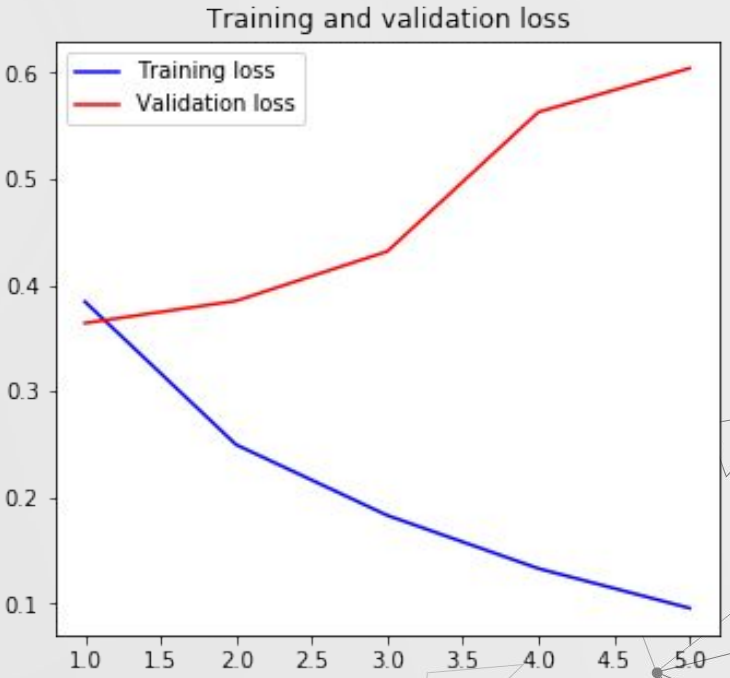
r/Anxiety
(classes imbalanced)



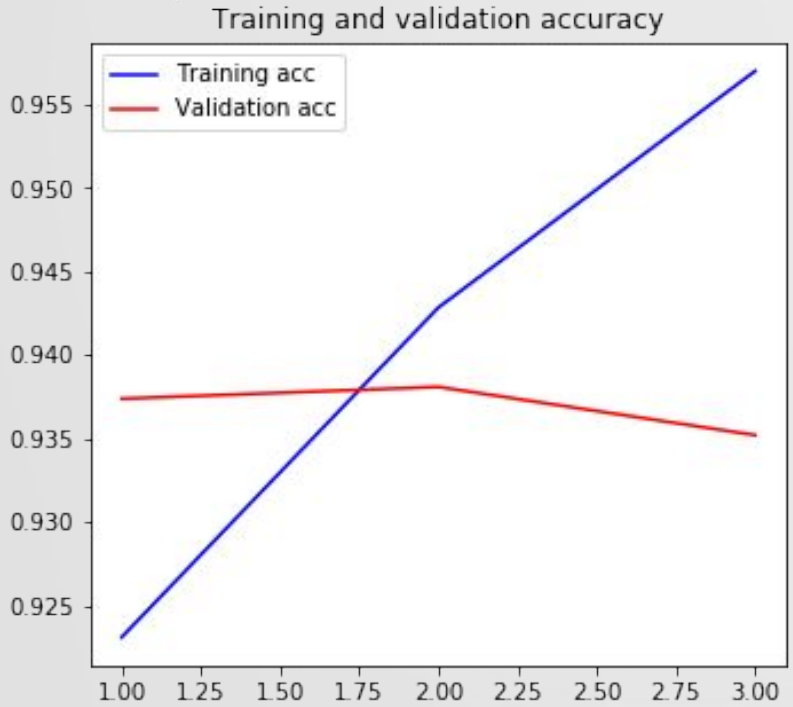


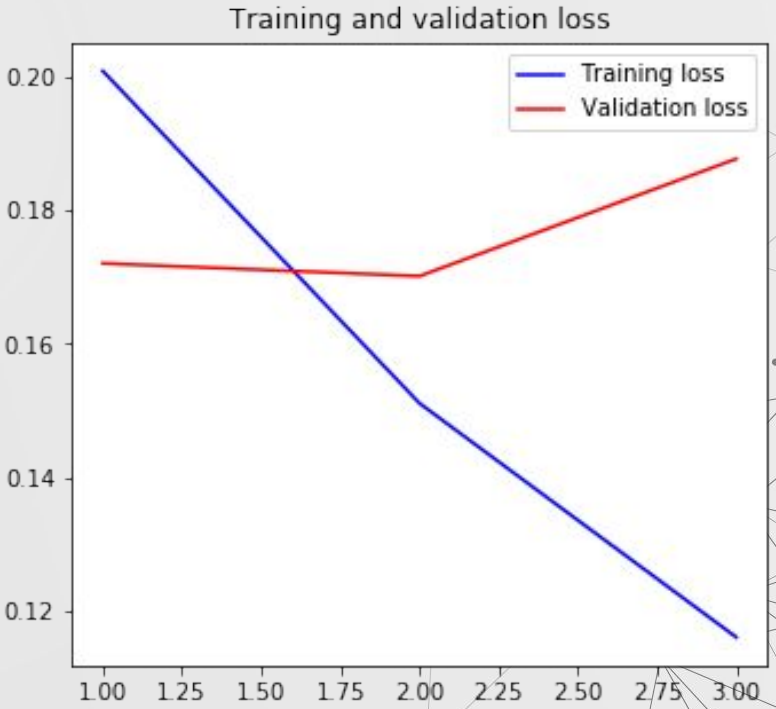
r/Anxiety
(classes balanced with SMOTE)



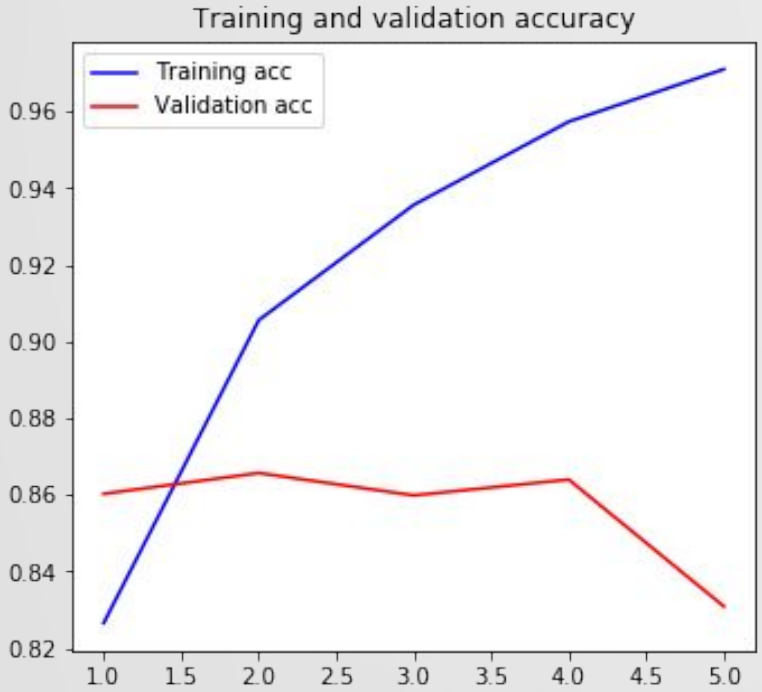


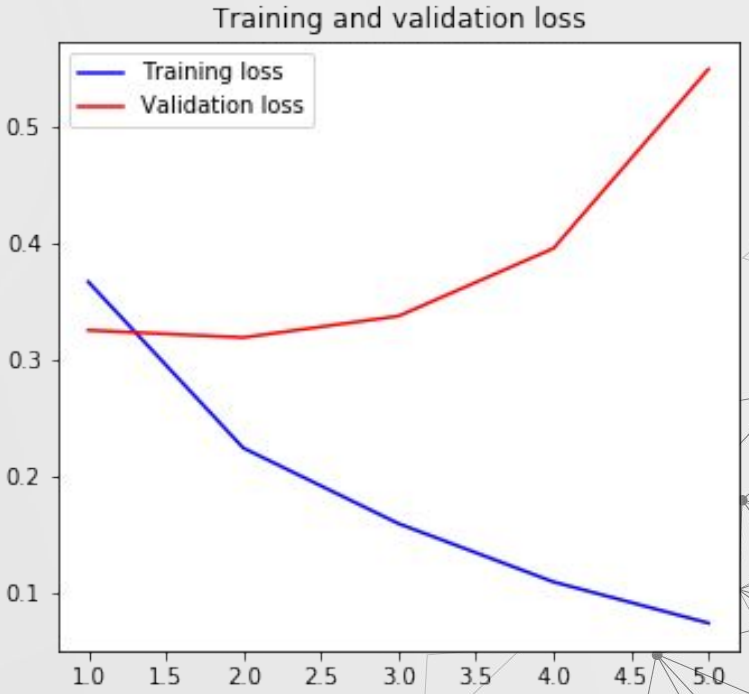
r/Schizophrenia
(classes imbalanced)



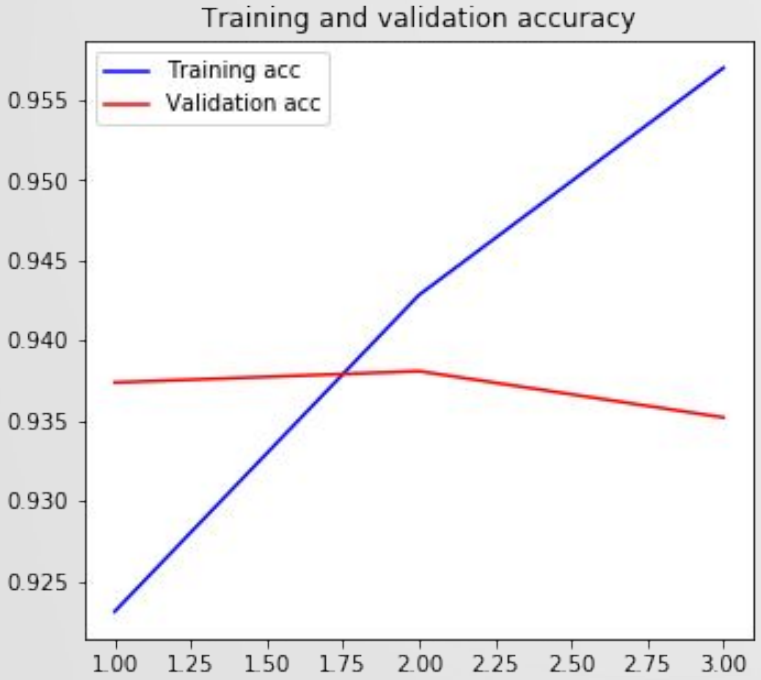


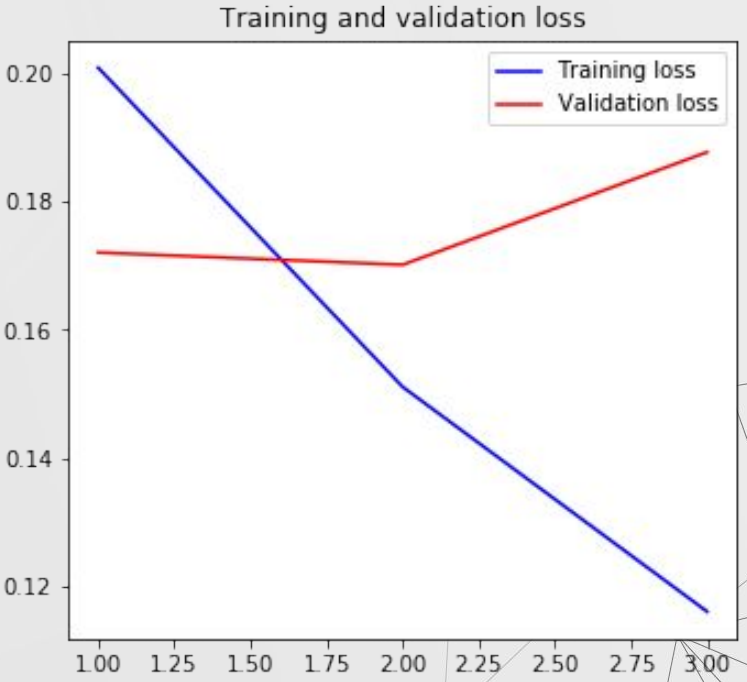
r/Schizophrenia
(classes balanced with SMOTE)



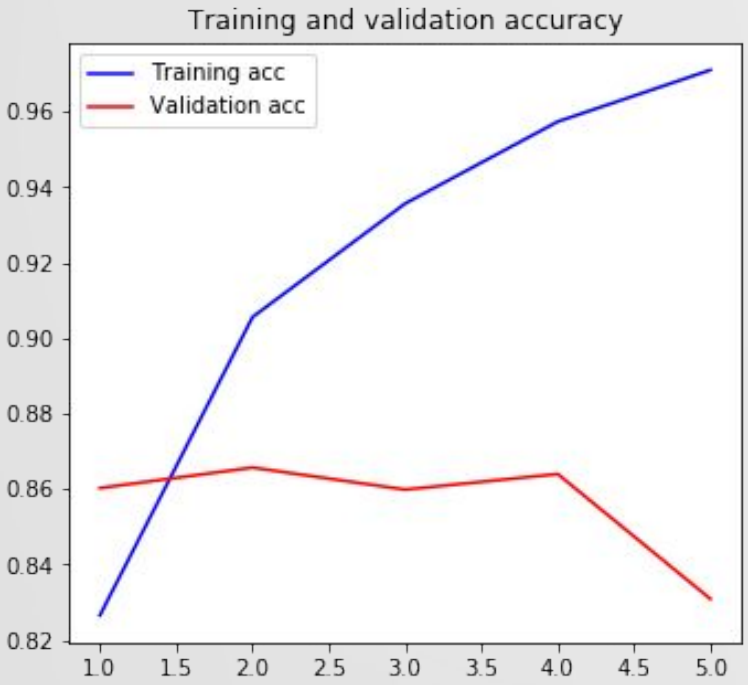


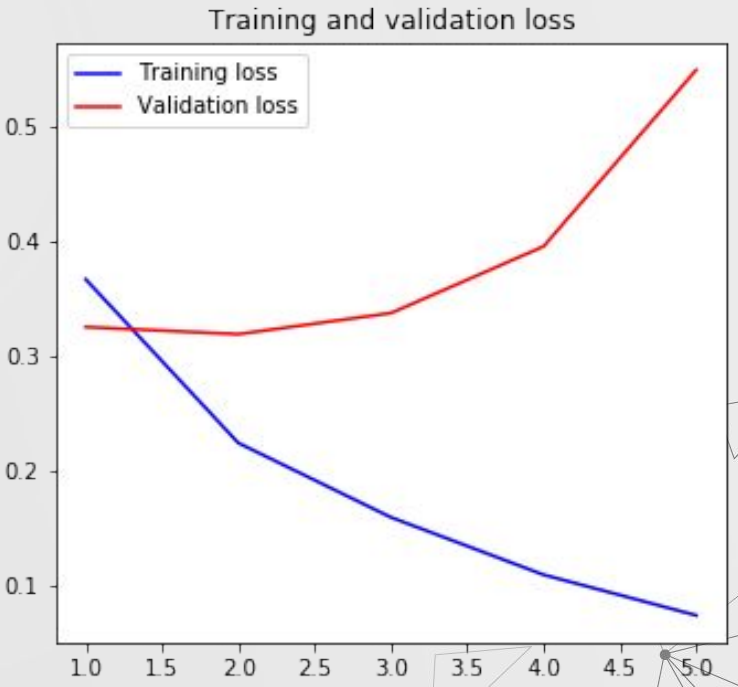
r/AnorexiaNervosa (classes imbalanced)





r/AnorexiaNervosa
(classes balanced with SMOTE)





05

Conclusions

- We can obtain high accuracy scores with our neural networks, however more tuning is needed to ensure we are not overfitting.

- Read papers with a healthy dose of skepticism.

- This is an important research area and more work is needed on the topic.

THANK YOU!

Questions?