R/DEPRESSION VS R/SHOWERTHOUGHTS

Is it a shower thought or is it worse?

Nader Esmael

Disclaimer

The contents within the presentation, such as text, graphics, images, and other material, may contain language that is NSFW, crude, or offensive. The material may also be very sensitive as it will go over topics regarding depression. If this is too much too handle, you are free to not participate. The content is intended to inform and is not a substitute for a qualified professional's diagnosis and/or expertise.

Content Synopsis

OVERVIEW OF KEY IDEAS

Background

Problem Statement

Preprocessing & Insights

Modeling & Metrics

Words of Importance

Conclusions & Improvements

Background

WHAT IS (CLINICAL) DEPRESSION?

It is a "mood disorder that causes distressing symptoms that affect how you feel, think, and handle daily activities, such as sleeping, eating, or working".

To be diagnosed with depression, these feelings and symptoms must be prevalent for 2 weeks consistently.

National Institute of Mental Health - https://www.nimh.nih.gov/health/publications/depression/index.shtml

Problem Statement

The Task

Utilize classification models to distinguish whether a post came from **r/depression** subreddit or **r/showerthoughts** subreddit.

The Selections

Posts from **r/depression** can range from a person's day to a person asking for help. Posts from **r/showerthoughts** has the randomness factor of what a thought of an average person can be.

The Comparison

To match the range of intensity with the range of randomness.

Example: "Negative numbers are actually bigger than positive numbers because they have a minus sign at the front."

Preprocessing

The Data

Scraped from the subreddits using Pushshift. 10,000 of the most recent posts from each subreddit was taken, concatenated, and put into a dataframe.

The Columns

Columns for word count and character count were added. The subreddit column was mapped where depression = 1 and shower thoughts = 0.

Two more columns were added where I Lemmatized and PorterStemmed the title posts myself. The "selftext" column was dropped from the dataframe.

Insights

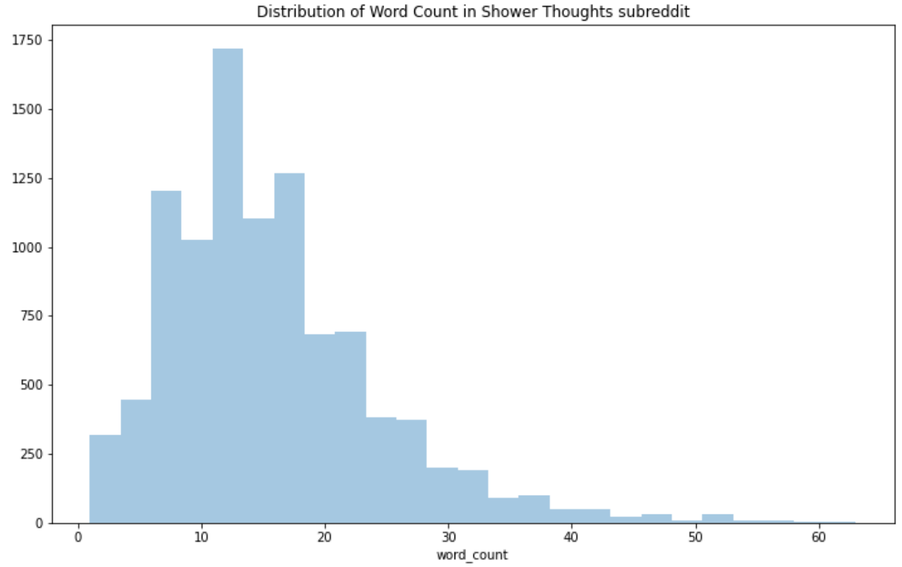
Top Titles	Count
"Help"	16
"."	13
"Alone"	10
"Tired"	8
"I'm so tired"	7
"Struggling"	7
"I don't know what to do"	7
"Depression"	7
":("	6
"I need help"	6

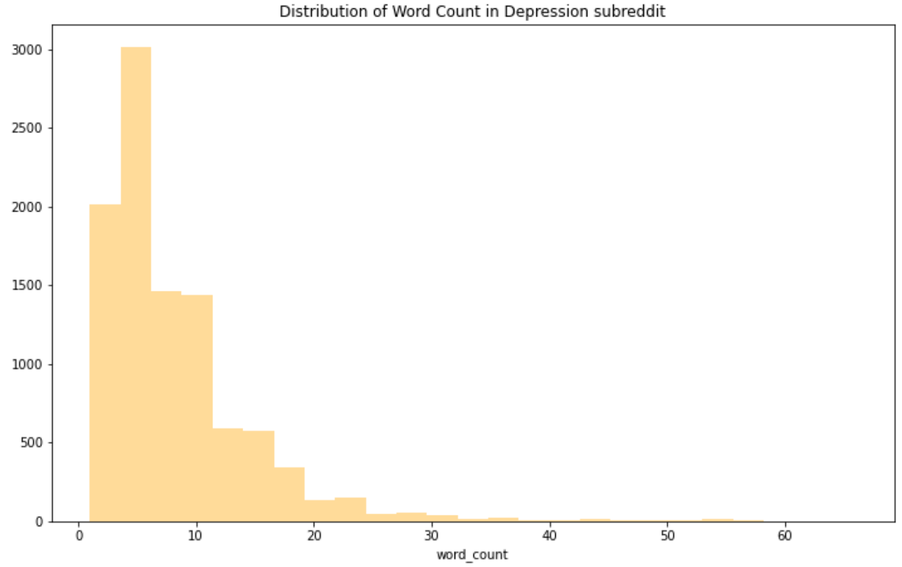
300 characters

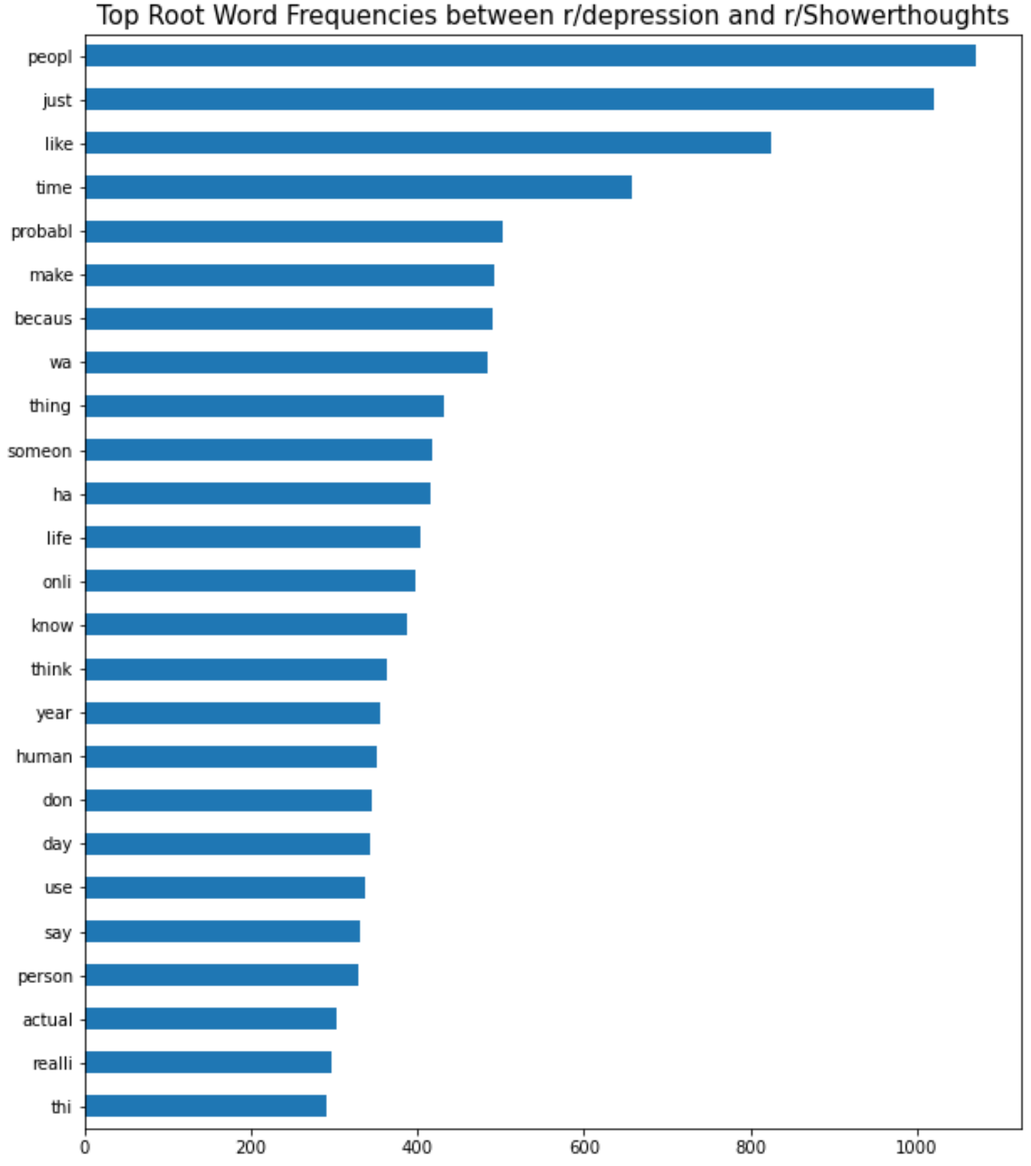
The limit to a title post. This could be a factor as to why r/depression encourages people to provide descriptions rather than just titles. **Another reason why I dropped the "selftext" column was because it would have made the model to obvious.**

9,593 / 9,979

This is the number of unique posts for r/depression and r/showerthoughts, respectively.


Distribution of Word Count in Shower Thoughts subreddit
1750
1500
1250
1000
750
500
250
0
0
10
20
30
40
50
60
word_count


Distribution of Word Count in Depression subreddit
3000
2500
2000
1500
1000
500
0
0
10
20
30
40
50
60
word_count


Top Root Word Frequencies between r/depression and r/Showerthoughts
peopl
just
like
time
probabl
make
becaus
wa
thing
someon
ha
life
onli
know
think
year
human
don
day
use
say
person
actual
realli
thi
0
200
400
600
800
1000

Mental pain is less dramatic than physical pain, but it is more common and also more hard to bear.

C.S. Lewis

Modeling & Results

(Pretty much the conclusion)

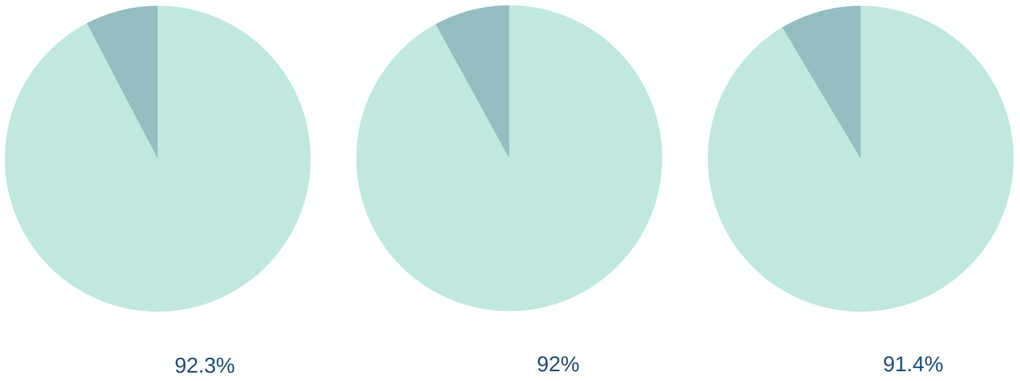
Classification Model & Hyper Parameters

Logistic Regression

TfidVectorizer

{'model__C': 1,
 'model__penalty': 'l2',
 'trans__max_df': 0.85,
 'trans__max_features': 3000,
 'trans__min_df': 5,
 'trans__ngram_range': (1, 3),
 'trans__stop_words': None}

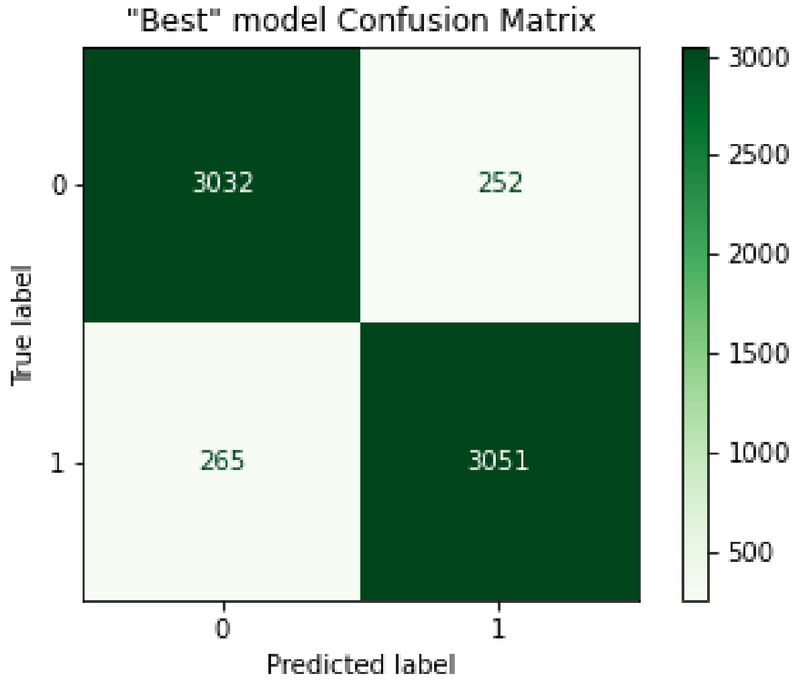
Classification Metrics & Confusion Matrix



Specificity

Sensitivity

Best Score/Accuracy



Words of Importance

COEFFICIENTS AND EXPONENTIATED COEFFICIENTS

Since I used a Logistic Regression as my best model, I looked at the coefficients and used that to determine whether a specific word added significance to a title being a **r/depression** post or not.

	word(s)	coefficient	exp_coef
592	depress	9.040511	8438.084964
1603	my	5.688897	295.567368
1517	me	5.258693	192.229999
819	feel	5.007686	149.558261
2254	suicid	3.840616	46.554127
1066	help	3.816397	45.440198
1630	myself	3.583370	35.994655
165	anyon	2.918018	18.504566
1233	is it	2.739656	15.481652
827	feel like	2.697899	14.848507
2725	want to	2.660107	14.297823
2856	wish	2.628148	13.848098
904	friend	2.547431	12.774250
2525	tire	2.481849	11.963363
1540	mental	2.472815	11.855772
1148	im	2.458320	11.685165
2035	sad	2.432114	11.382917
2443	therapi	2.426240	11.316249
1021	happi	2.308422	10.058536
56	advic	2.259658	9.579815
2237	struggl	2.173840	8.791984
916	fuck	2.138048	8.482860
2699	vent	2.137042	8.474333
2724	want	2.120504	8.335341
1029	hate	2.113494	8.277107

r/Showerthoughts vs. r/depression

GUESS WHICH SUBREDDIT IT CAME FROM?

'Forget a person you considered everything'

'I give advice to people to make decisions that I am not brave enough to make'

r/Showerthoughts vs. r/depression

GUESS WHICH SUBREDDIT IT CAME FROM?

MISCLASSIFIED

Depression

'Forget a person you considered everything'

Shower Thought

'I give advice to people to make decisions that I am not brave enough to make'

Probable Improvements

Obviously better hyper parameters, but also...

A SENTIMENT ANALYZER

Grasp a better understanding of this analyzer to correlate coefficients with sentiment score.

MORE STOP WORDS

Create a better dictionary for stop words that using the defaults from Python.

RESEARCH MORE

Get a better understanding of depression to optimize the model.

A BETTER MODEL

Utilize the other modeling techniques such as bootstrapping or boosting.

Is it applicable?

IF THE MODEL WERE PERFECTED...

It could be used on other social media platforms and to detect if the user is suffering from depression or having thoughts of depression.




Thank you!
Questions?