NATIONAL SUICIDE PREVENTION LIFELINE™
1-800-273-TALK
www.suicidepreventionlifeline.org

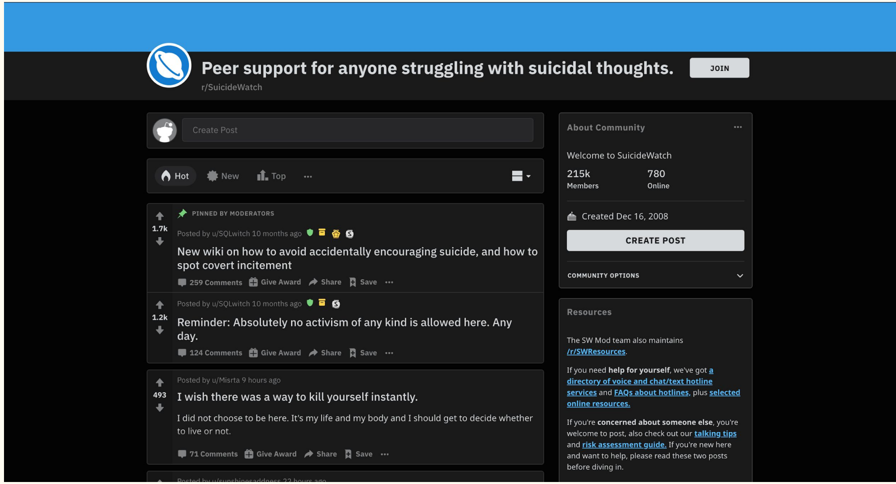
Eu Jin Lee - Data Science Immersive - Project 3

Subreddit Profiler: LifeLine for NSPL

Classification of Subreddit Posts With Natural Language Processing

Mental Health is Real.

- r/SuicideWatch and r/OffMyChest: two subreddit support groups that allow redditors a safe space to express themselves, their concerns and receive possibly life saving advice.
- Noticeably from the titles of the subreddits, one is a lot heavier than the other.
- Reddit is a common space for younger generation.
- The goal of the National Suicide Prevention is to monitor these subreddits for users' posts that suggest self harm and notify our response teams.



Peer support for anyone struggling with suicidal thoughts.

r/SuicideWatch

PINNED BY MODERATORS

1.7k

Posted by u/SQLwitch 10 months ago

New wiki on how to avoid accidentally encouraging suicide, and how to spot covert incitement

259 Comments

1.2k

Posted by u/SQLwitch 10 months ago

Reminder: Absolutely no activism of any kind is allowed here. Any day.

124 Comments

493

Posted by u/Misrta 9 hours ago

I wish there was a way to kill yourself instantly.

I did not choose to be here. It's my life and my body and I should get to decide whether to live or not.

71 Comments

Posted by u/sunshinesaddness 22 hours ago

About Community

Welcome to SuicideWatch

215k Members

780 Online

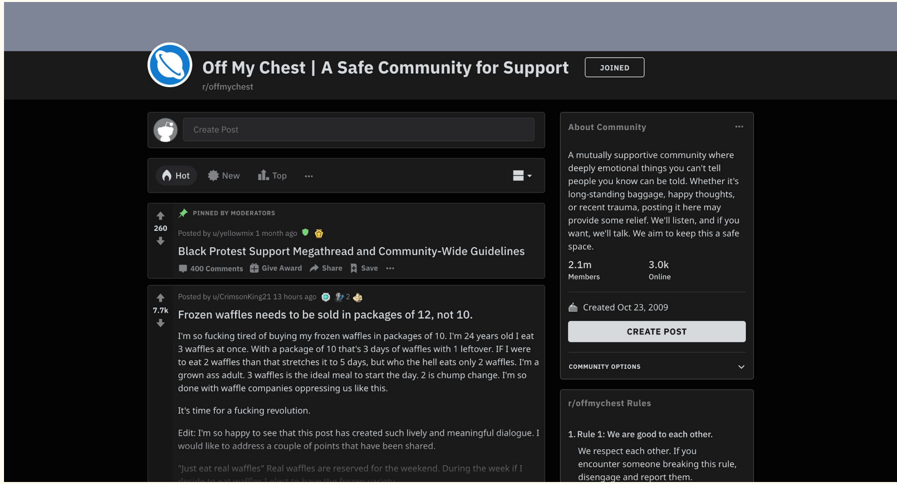
Created Dec 16, 2008

Resources

The SW Mod team also maintains /r/SWResources.

If you need **help for yourself**, we've got a directory of voice and chat/text hotline services and FAQs about hotlines, plus selected online resources.

If you're **concerned about someone else**, you're welcome to post, also check out our talking tips and risk assessment guide. If you're new here and want to help, please read these two posts before diving in.



Off My Chest | A Safe Community for Support

r/offmychest

PINNED BY MODERATORS

260

Posted by u/yellowmix 1 month ago

Black Protest Support Megathread and Community-Wide Guidelines

400 Comments

7.7k

Posted by u/CrimsonKing21 13 hours ago 2

Frozen waffles needs to be sold in packages of 12, not 10.

I'm so fucking tired of buying my frozen waffles in packages of 10. I'm 24 years old I eat 3 waffles at once. With a package of 10 that's 3 days of waffles with 1 leftover. IF I were to eat 2 waffles than that stretches it to 5 days, but who the hell eats only 2 waffles. I'm a grown ass adult. 3 waffles is the ideal meal to start the day. 2 is chump change. I'm so done with waffle companies oppressing us like this.

It's time for a fucking revolution.

Edit: I'm so happy to see that this post has created such lively and meaningful dialogue. I would like to address a couple of points that have been shared.

"Just eat real waffles" Real waffles are reserved for the weekend. During the week if I decide to eat waffles I elect to have the frozen variety

About Community

A mutually supportive community where deeply emotional things you can't tell people you know can be told. Whether it's long-standing baggage, happy thoughts, or recent trauma, posting it here may provide some relief. We'll listen, and if you want, we'll talk. We aim to keep this a safe space.

2.1m Members

3.0k Online

Created Oct 23, 2009

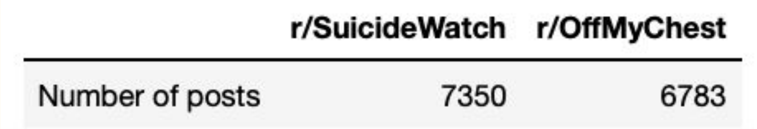
r/offmychest Rules

1. Rule 1: We are good to each other.

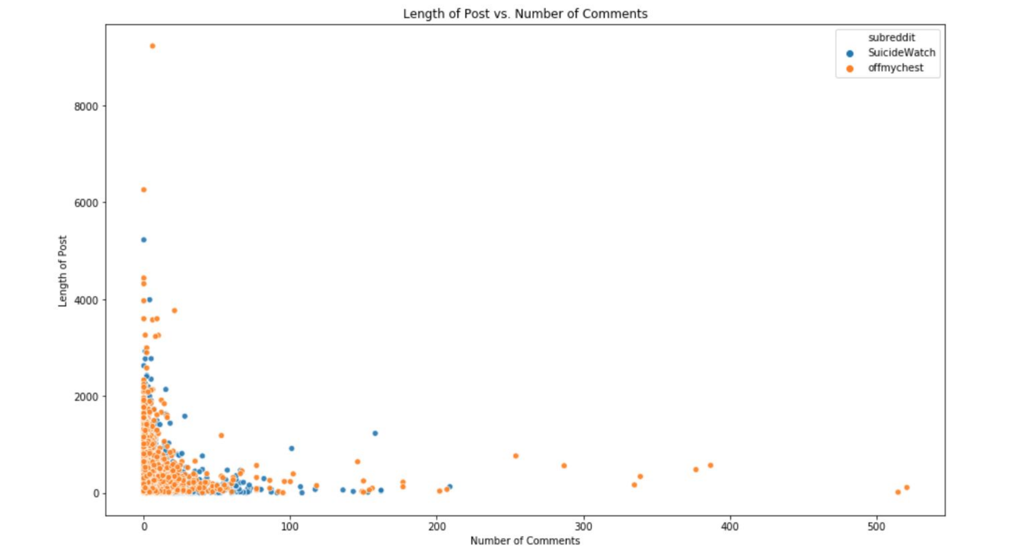
We respect each other. If you encounter someone breaking this rule, disengage and report them.

Data Cleaning and Processing

- Pushshift is an open-source alternative to Reddit's API that allows us to pull data going back in time, and returns data in a more convenient format.
- Collected first 1000 posts from each subreddit at 12-day intervals, going back two years
- Cleaned away duplicates, moderator boilerplate, [deleted], [removed] type posts.
- Final clean dataset contains ~ 14133 posts from both subreddits.



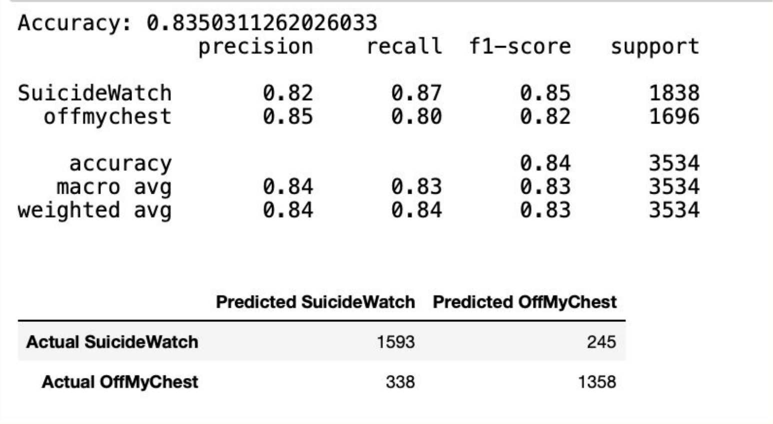
	r/SuicideWatch	r/OffMyChest
Number of posts	7350	6783



Length of Post vs. Number of Comments

Modeling using Logistic Regression

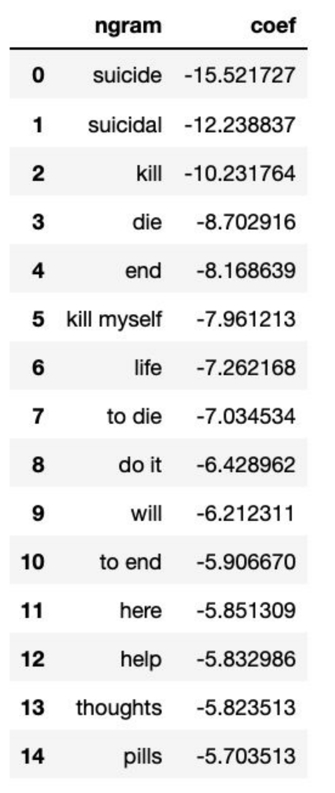
- Logistic Regression is a classic technique used in large scale classification problems, that has the advantage of being highly interpretable which is important to us.
- We looked at the word pairings or n-grams the model associates most with self harm statements. We can also observe the intent of these posts and classify them back to their original subreddits. In future we hope to develop a triage system to classify the level of intent and be able to act accordingly.
- We use GridSearch to look at different combinations of vectorizer options and Logistic Regression options, and find the best performing fit.
- On new data, our model predicts the correct subreddit 83.5% of the time (baseline accuracy was 52 %!) Outperforming MultinomialNB and SVM models.
- The confusion matrix on the (right) shows that SuicideWatch posts were slightly easier to identify.



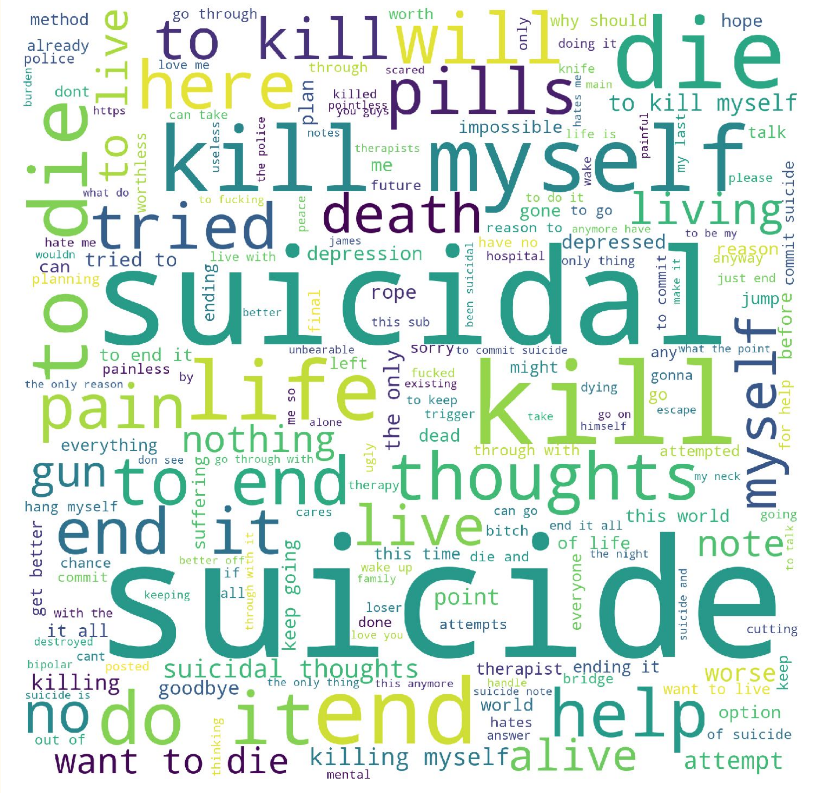
```
Accuracy: 0.8350311262026033
              precision    recall  f1-score   support

SuicideWatch       0.82      0.87      0.85      1838
  offmychest       0.85      0.80      0.82      1696

    accuracy                           0.84      3534
   macro avg       0.84      0.83      0.83      3534
weighted avg       0.84      0.84      0.83      3534
```

	Predicted SuicideWatch	Predicted OffMyChest
Actual SuicideWatch	1593	245
Actual OffMyChest	338	1358

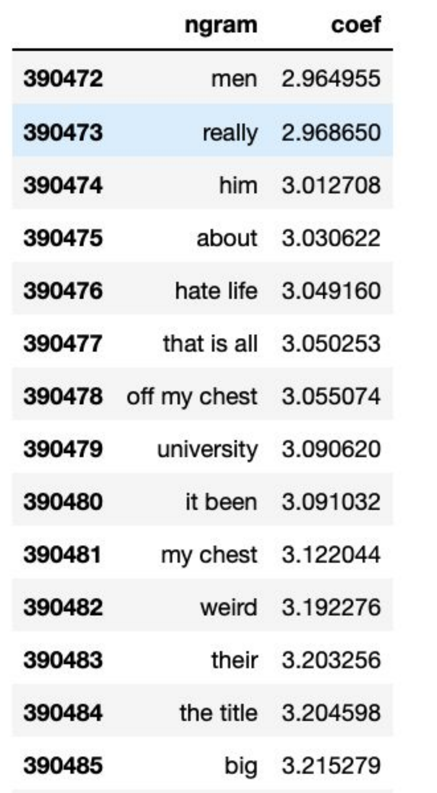
R/SuicideWatch



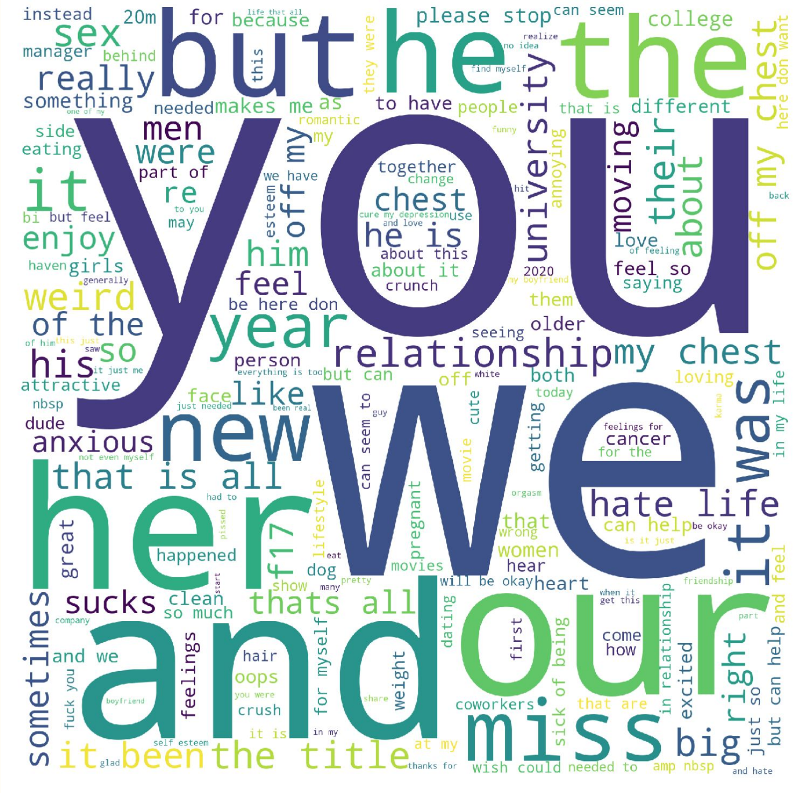
	ngram	coef
0	suicide	-15.521727
1	suicidal	-12.238837
2	kill	-10.231764
3	die	-8.702916
4	end	-8.168639
5	kill myself	-7.961213
6	life	-7.262168
7	to die	-7.034534
8	do it	-6.428962
9	will	-6.212311
10	to end	-5.906670
11	here	-5.851309
12	help	-5.832986
13	thoughts	-5.823513
14	pills	-5.703513



R/OffMyChest



	ngram	coef
390472	men	2.964955
390473	really	2.968650
390474	him	3.012708
390475	about	3.030622
390476	hate life	3.049160
390477	that is all	3.050253
390478	off my chest	3.055074
390479	university	3.090620
390480	it been	3.091032
390481	my chest	3.122044
390482	weird	3.192276
390483	their	3.203256
390484	the title	3.204598
390485	big	3.215279



Conclusion and Recommendations

- Dive deeper into the word frequencies and develop narratives and theories that may be the most common stressors for people. Find possible solutions.
- Be aware of the fact that the demographic on Reddit is not a complete representation of the population suffering from mental health issues but it is a good step in the right direction.

Moving Forward

- Feature engineer a ranking system that is able to quantify and rank the severity of intent within a post and provide a support system accordingly.

If you or a loved one is thinking about suicide or would like emotional support, the Lifeline network is available 24/7 across the United States.

1-800-273-8255

https://suicidepreventionlifeline.org/