How well does a Classification model hold up over time?

Aida Rahim

3/5/2021

Questions to Address

- Is it possible to differentiate posts by subreddit? Criteria:
  - Similar content
  - Huge membership
  - Regular, non-specialized daily language
- Which model is best?
- How do the top words change over time?
- Does prediction accuracy change over time?
- What is the model actually classifying?

reddit

r/LifeProTips (19M)

Tips that improve your life in one way or another

LPT: If you cannot afford waterproof footwear, place your feet in a gallon food storage bag (one bag per foot, over top of socks) and then put on your shoes as normal. The plastic barrier is thin enough for your shoe to still fit properly while keeping your toes dry.

If you ever want to mark the end of your tape, just use a paper clip!

VS

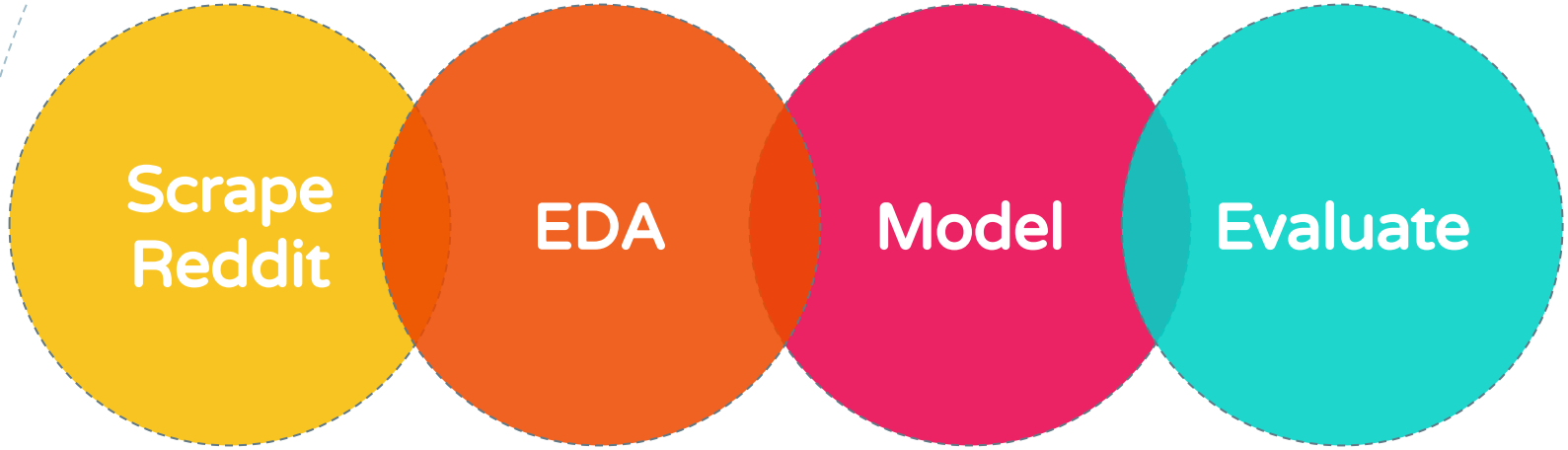
r/Showerthoughts (22M)

For sharing those miniature epiphanies you have that highlight the oddities within the familiar

What we see as terrifying, might be a beautiful piece of art for an alien from another world

At all points in time you can hear both outside and inside your own mouth

PROCESS



pushshift API

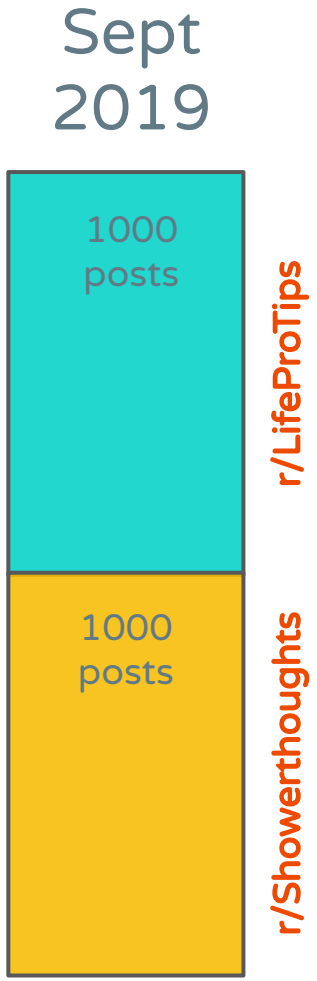
Logistic Regression
Random Forest
ExtraTrees
Support Vector Machine

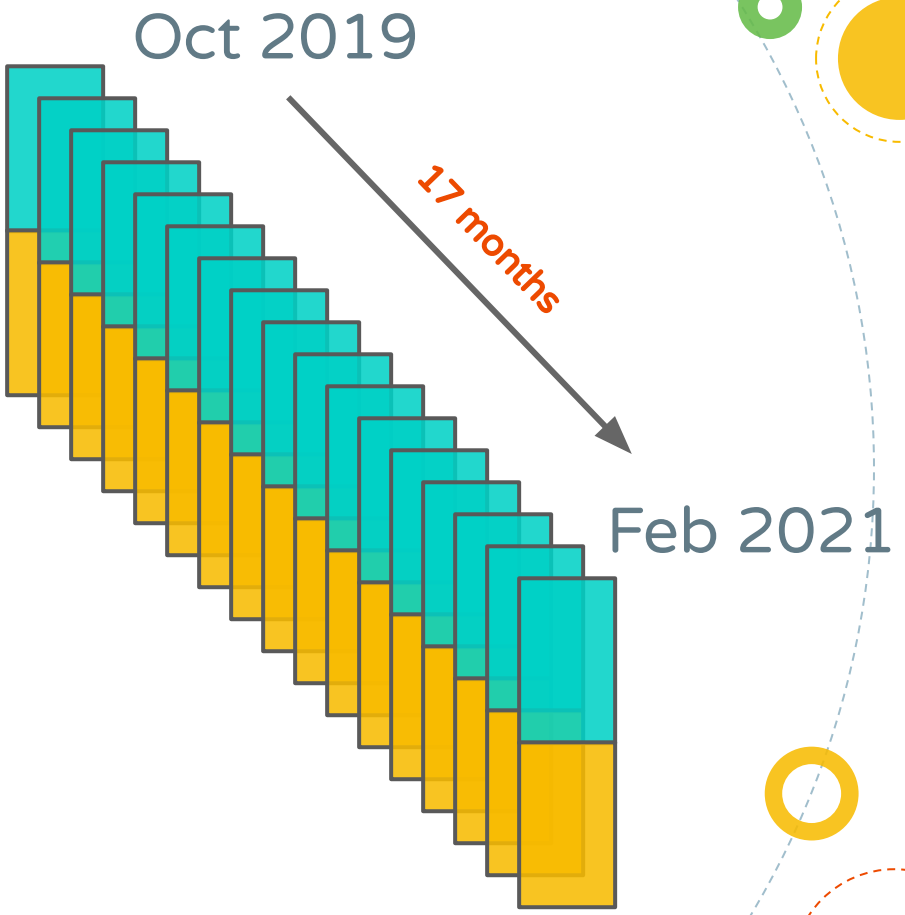
Scrape Reddit

Scrape Reddit

1000 posts = 6 days

Starting midnight on the 20th of each month



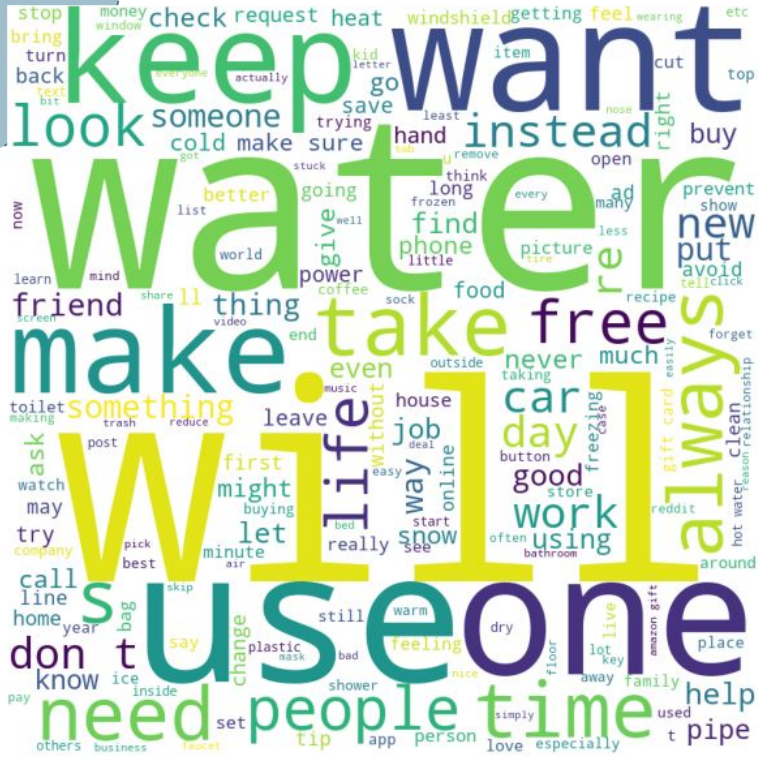


Exploratory Data Analysis

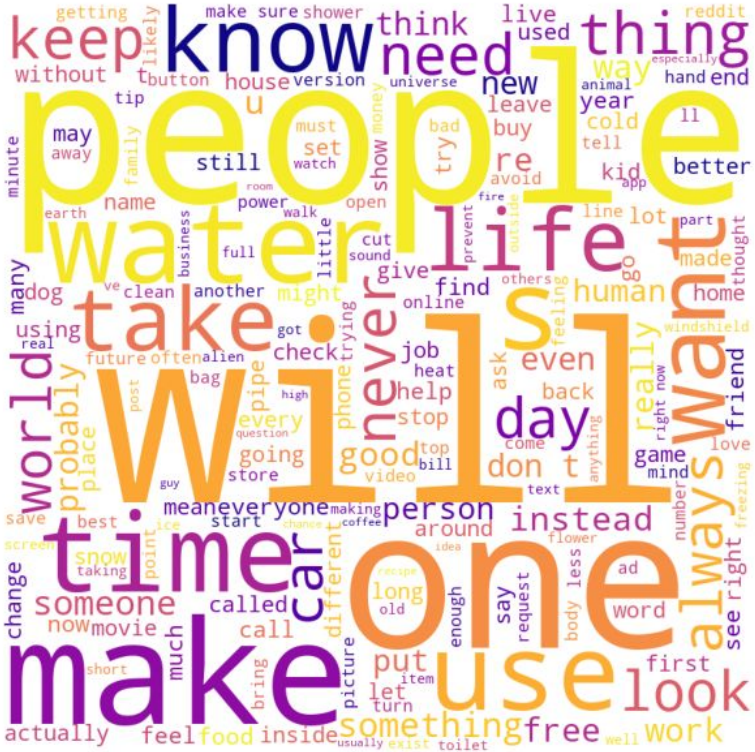
EDA

Subreddit Flavors

r/LifeProTips (1)

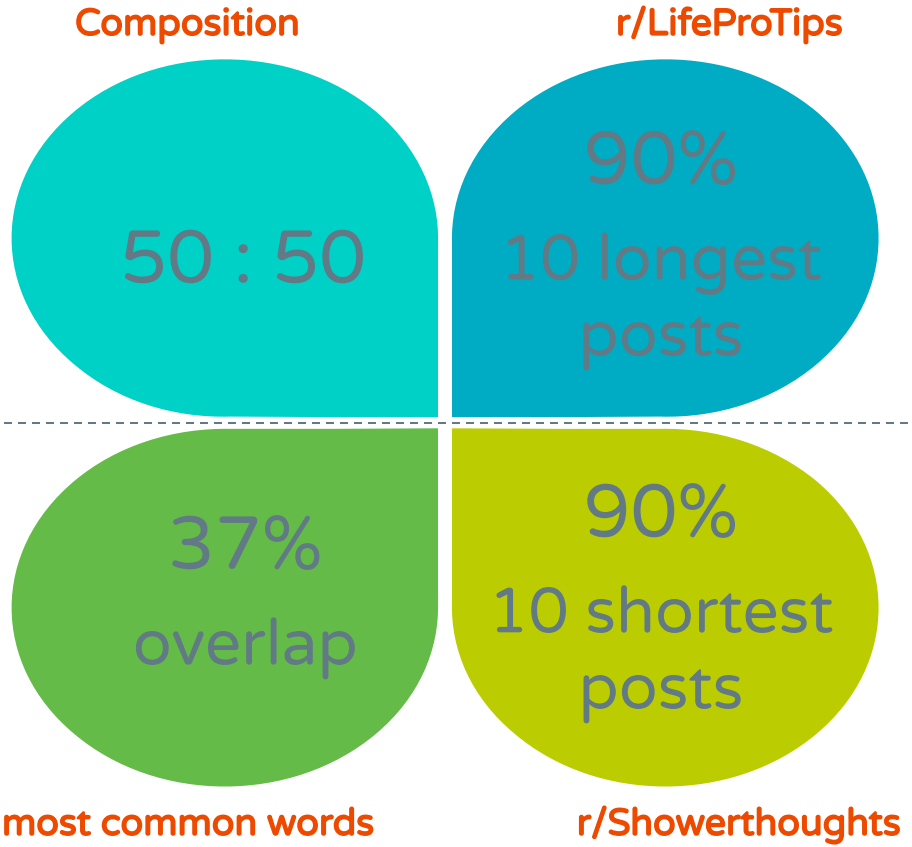


(0) r/Showerthoughts



EDA

Subreddit Flavors

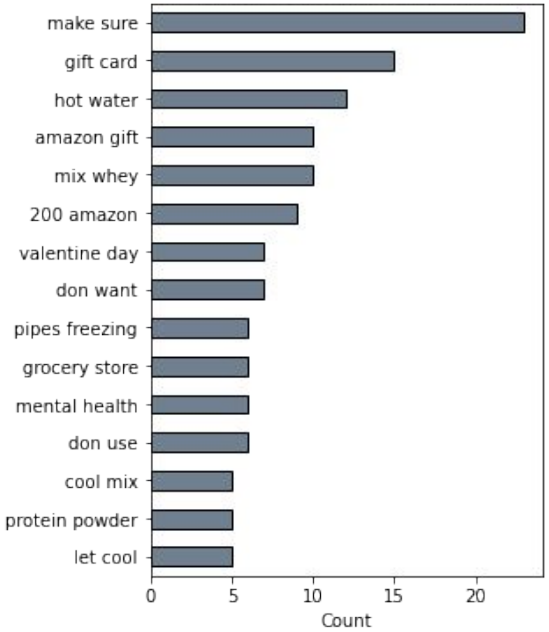


EDA

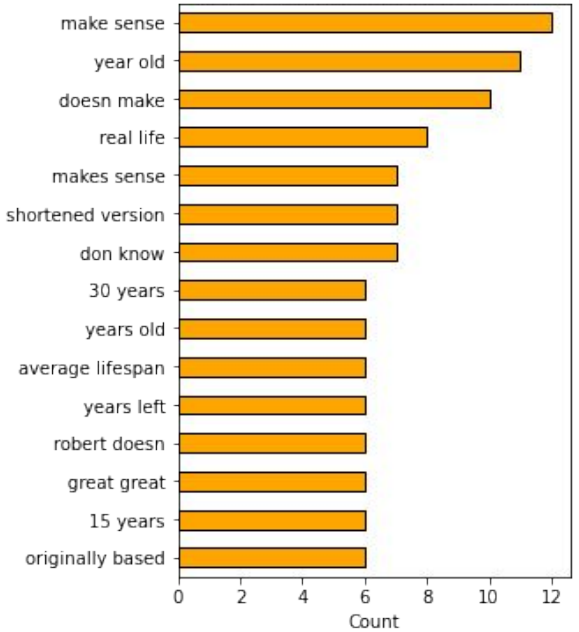
Subreddit Flavors

WORDS, WORDS, WORDS
TOP 15 BIGRAMS

r/LifeProTips



r/Showerthoughts



Practical

Deep

Build Model

Model

74.5% of posts from the 'LifeProTips' contain the identifier 'LPT'

Model	Transformer	Estimator	Details	Accuracy (train)	Accuracy (test)
1	Tfidf	LogReg	'lpt' stopword	0.926	0.852
2	Tfidf	LogReg	No pipe, no stopwords	0.955	0.903
3	CountVect	LogReg	Pipe, GridSearchCV, 'lpt' stopword	0.988	0.836
4	Tfidf	LogReg		0.917	0.844
5	Tfidf	LogRegCV		0.951	0.84
6	Tfidf	RandomForest		0.881	0.809
7	Tfidf	ExtraTrees		0.903	0.822
8	Tfidf	SVM		0.991	0.854

Model

Select 2 Models for Performance Comparison on Testing Data

Model	Transformer	Estimator	Details	Accuracy (train)	Accuracy (test)
1	Tfidf	LogReg	'lpt' stopword	0.926	0.852
2	Tfidf	LogReg	No pipe, no stopwords	0.955	0.903
3	CountVect	LogReg	Pipe, GridSearchCV, 'lpt' stopword	0.988	0.836
4	Tfidf	LogReg		0.917	0.844
5	Tfidf	LogRegCV		0.951	0.84
6	Tfidf	RandomForest		0.881	0.809
7	Tfidf	ExtraTrees		0.903	0.822
8	Tfidf	SVM		0.991	0.854

Model

Logistic Regression Inference

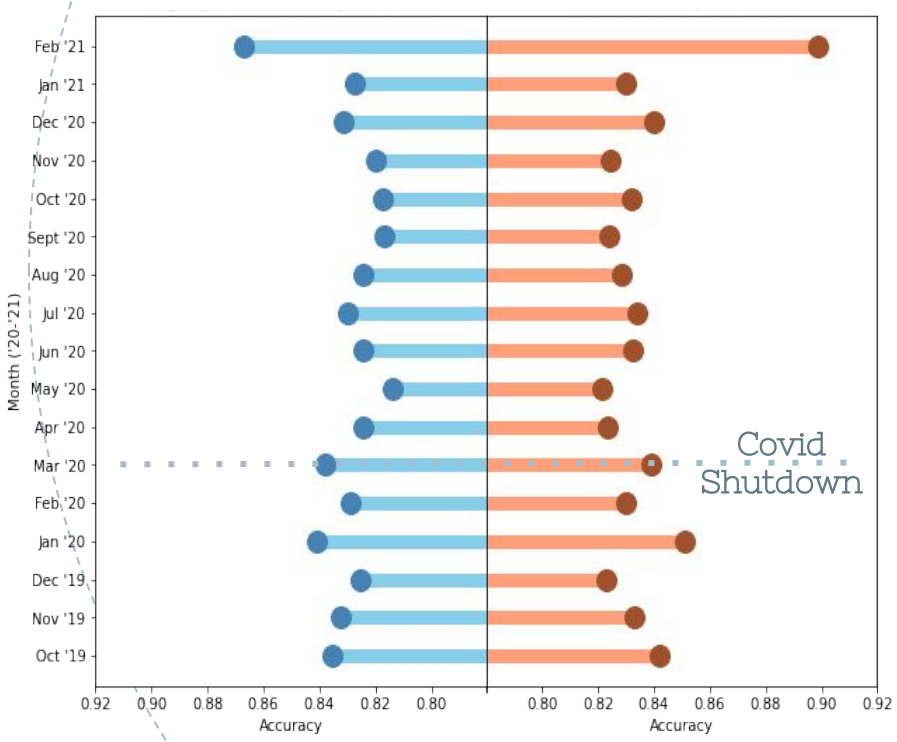
	Coefficient
your	126.027519
don	13.146453
use	9.150317
water	6.384216
for	6.292872
keep	6.159003
you	5.308579
want	5.239847
and	4.890368
to	4.889101

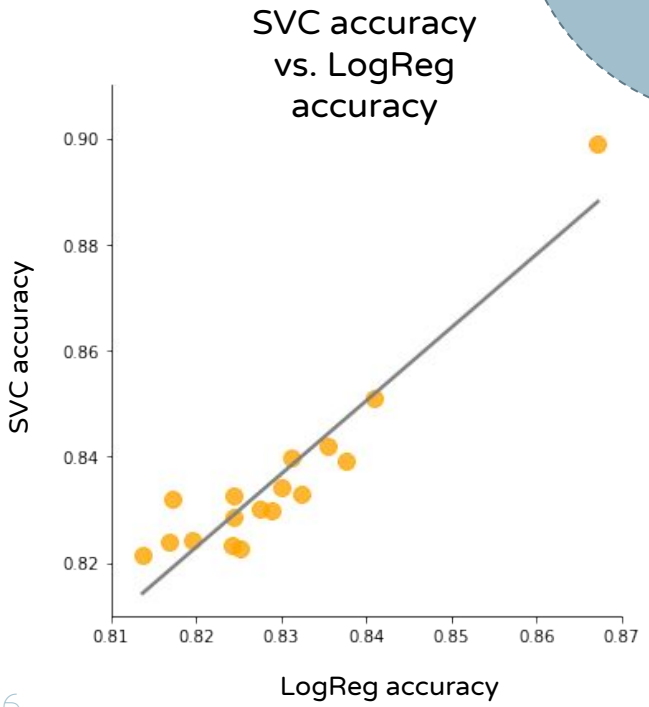
- Each increase in ‘your’ makes a sentence 126 times more likely to be from LifeProTips (all else remaining equal)
- ‘Your’ is 10 times more important than the next most important word
- Misclassification rate:
  - 53% LifeProTips : 47% Showerthoughts
- Analysis of misclassified posts
  - 17% overlap in 100 most common words
  - Nothing else stands out

Apply 2 Models to 17 Test Datasets

Model Accuracies Over Time

Evaluate





Top 15 Words Over Time

r/LifeProTips

	Oct '19	Nov '19	Dec '19	Jan '20	Feb '20	Mar '20	Apr '20	May '20	Jun '20	Jul '20	Aug '20	Sept '20	Oct '20	Nov '20	Dec '20	Jan '21	Feb '21
0	don	d8	don	don	don	time	don	don	don	don	don	don	don	don	don	don	don
1	use	d9	make	use	use	don	use	just	like	use	use	want	make	make	time	use	water
2	want	don	just	want	just	use	time	make	use	want	want	like	time	use	want	make	use
3	make	just	want	make	want	home	want	use	make	time	time	use	want	time	make	time	make
4	time	want	use	like	time	paper	like	want	time	make	make	make	use	just	just	just	just
5	phone	use	time	just	people	toilet	make	time	just	phone	like	need	just	people	use	people	want
6	just	phone	know	phone	phone	people	just	like	want	like	way	time	phone	good	need	want	time
7	d9	make	christmas	car	need	like	need	people	need	just	just	just	way	like	people	like	need
8	like	like	people	instead	make	make	try	instead	people	know	people	day	like	try	like	life	free
9	ll	time	instead	new	like	covid	phone	way	phone	need	know	people	need	want	ask	phone	people
10	d8	a7	like	time	way	quarantine	free	ll	instead	people	instead	new	people	need	way	ll	like
11	need	try	ll	try	stop	just	check	try	feel	ask	life	life	life	know	new	need	life
12	way	b1	save	water	try	19	save	know	ask	person	phone	phone	ll	ll	help	way	car
13	new	ll	new	check	day	need	car	ask	life	life	need	try	request	phone	know	know	look
14	car	need	water	ask	help	want	new	need	know	new	new	way	know	instead	instead	free	instead

r/Showerthoughts

	Oct '19	Nov '19	Dec '19	Jan '20	Feb '20	Mar '20	Apr '20	May '20	Jun '20	Jul '20	Aug '20	Sept '20	Oct '20	Nov '20	Dec '20	Jan '21	Feb '21
0	people	people	people	just	people	people	people	people	just	people	just	people	people	people	people	people	people
1	just	just	just	people	just	just	just	just	people	just	people	just	just	just	just	just	just
2	like	like	like	like	like	time	like	like	like	like	like	like	like	like	like	like	like
3	probably	time	probably	play	time	probably	time	probably	time	probably	don	life	probably	time	time	life	time
4	time	probably	time	probably	life	world	probably	time	probably	life	time	probably	time	don	probably	time	don
5	don	don	good	life	don	like	don	make	really	time	probably	time	life	probably	life	don	life
6	world	think	life	time	probably	going	life	world	make	world	know	know	make	life	really	probably	make
7	know	day	person	man	day	covid	world	don	know	don	really	make	know	really	actually	make	mars
8	think	world	don	know	make	coronavirus	person	know	don	make	life	don	way	make	new	world	probably
9	person	make	way	don	know	right	say	person	day	think	world	good	world	think	know	think	know
10	day	life	say	day	think	quarantine	make	life	life	actually	make	day	don	person	don	new	world
11	years	know	look	think	really	virus	human	ve	actually	know	new	bad	really	way	year	years	things
12	life	good	really	make	say	19	think	right	water	water	day	really	technically	water	think	day	think
13	called	having	know	water	different	home	going	lot	say	years	person	person	water	know	way	human	earth
14	water	lot	world	actually	world	think	really	really	years	person	water	actually	human	actually	ve	know	going

Evaluate

Top 15 Words Over Time

r/LifeProTips

	Oct '19	Nov '19	Dec '19	Jan '20	Feb '20	Mar '20	Apr '20	May '20	Jun '20	Jul '20	Aug '20	Sept '20	Oct '20	Nov '20	Dec '20	Jan '21	Feb '21
0	don	d8	don	don	don	time	don	don	don	don	don	don	don	don	don	don	don
1	use	d9	make	use	use	don	use	just	like	use	use	want	make	make	time	use	water
2	want	don	just	want	just	use	time	make	use	want	want	like	time	use	want	make	use
3	make	just	want	make	want	home	want	use	make	time	time	use	want	time	make	time	make
4	time	want	use	like	time	paper	like	want	time	make	make	make	use	just	just	just	just
5	phone	use	time	just	people	toilet	make	time	just	phone	like	need	just	people	use	people	want
6					hone	people	just	like	want	like	way	time	phone	good	need	want	time
7					need	like	need	people	need	just	just	just	way	like	people	like	need
8					make	make	try	instead	people	know	people	day	like	try	like	life	free
9	ll	time	instead	new	like	covid	phone	way	phone	need	know	people	need	want	ask	phone	people
10	d8	a7	like	time	way	quarantine	free	ll	instead	people	instead	new	people	need	way	ll	like
11	need	try	ll	try	stop	just	check	try	feel	ask	life	life	life	know	new	need	life
12	way	b1	save	water	try	19	save	know	ask	person	phone	phone	ll	ll	help	way	car
13	new	ll	new	check	day	need	car	ask	life	life	need	try	request	phone	know	know	look
14	car	need	water	ask	help	want	new	need	know	new	new	way	know	instead	instead	free	instead

christmas

r/Showerthoughts

	Oct '19	Nov '19	Dec '19	Jan '20	Feb '20	Mar '20	Apr '20	May '20	Jun '20	Jul '20	Aug '20	Sept '20	Oct '20	Nov '20	Dec '20	Jan '21	Feb '21
0	people	people	people	just	people	people	people	people	just	people	just	people	people	people	people	people	people
1	just	just	just	people	just	just	just	just	people	just	people	just	just	just	just	just	just
2	like	like	like	like	like	time	like	like	like	like	like	like	like	like	like	like	like
3	probably	time	probably	play	time	probably	time	probably	time	probably	don	life	probably	time	time	life	time
4	time	probably	time	probably	life	world	probably	time	probably	life	time	probably	time	don	probably	time	don
5	don	don	good	life	don	like	don	make	really	time	probably	time	life	probably	life	don	life
6	world	think	life	time	probably	going	life	world	make	world	know	know	make	life	really	prob	
7	know	day	person	man	day	covid	world	don	know	don	really	make	know	really	actually		
8	think	world	don	know	make	coronavirus	person	know	don	make	life	don	way	make	new		
9	person	make	way	don	know	right	say	person	day	think	world	good	world	think	know	think	know
10	day	life	say	day	think	quarantine	make	life	life	actually	make	day	don	person	don	new	world
11	years	know	look	think	really	virus	human	ve	actually	know	new	bad	really	way	year	years	things
12	life	good	really	make	say	19	think	right	water	water	day	really	technically	water	think	day	think
13	called	having	know	water	different	home	going	lot	say	years	person	person	water	know	way	human	earth
14	water	lot	world	actually	world	think	really	really	years	person	water	actually	human	actually	ve	know	going

mars

Evaluate

Top 15 Words Over Time

r/LifeProTips

	Oct '19	Nov '19	Dec '19	Jan '20	Feb '20	Mar '20	Apr '20	May '20	Jun '20	Jul '20	Aug '20	Sept '20	Oct '20	Nov '20	Dec '20	Jan '21	Feb '21
0	don	d8	don	don	don	time	don	don	don	don	don	don	don	don	don	don	don
1	use	d9	make	use					like	use	use	want	make	make	time	use	water
2	want	don	just	want					use	want	want	like	time	use	want	make	use
3	make	just	want	make					make	time	time	use	want	time	make	time	make
4	time	want	use	like					time	make	make	make	use	just	just	just	just
5	phone	use	time	just					just	phone	like	need	just	people	use	people	want
6	just	phone	know	phone					want	like	way	time	phone	good	need	want	time
7	d9	make	christmas	ca					need	just	just	just	way	like	people	like	need
8	like	like	people	instead					people	know	people	day	like	try	like	life	free
9	ll	time	instead	new					phone	need	know	people	need	want	ask	phone	people
10	d8	a7	like	time					instead	people	instead	new	people	need	way	ll	like
11	need	try	ll	try					feel	ask	life	life	life	know	new	need	life
12	way	b1	save	water					ask	person	phone	phone	ll	ll	help	way	car
13	new	ll	new	check					life	life	need	try	request	phone	know	know	look
14	car	need	water	ask	help	want	new	need	know	new	new	way	know	instead	instead	free	instead

paper

toilet

covid

quarantine

19

r/Showerthoughts

	Oct '19	Nov '19	Dec '19	Jan '20	Feb '20	Mar '20	Apr '20	May '20	Jun '20	Jul '20	Aug '20	Sept '20	Oct '20	Nov '20	Dec '20	Jan '21	Feb '21
0	people	people	people	just	people	people	people	people	just	people	just	people	people	people	people	people	people
1	just	just	just	people					people	just	people	just	just	just	just	just	just
2	like	like	like	like					like	like	like	like	like	like	like	like	like
3	probably	time	probably	play					time	probably	don	life	probably	time	time	life	time
4	time	probably	time	probably					probably	life	time	probably	time	don	probably	time	don
5	don	don	good	life					really	time	probably	time	life	probably	life	don	life
6	world	think	life	time					make	world	know	know	make	life	really	probably	make
7	know	day	person	man					know	don	really	make	know	really	actually	make	mars
8	think	world	don	know					don	make	life	don	way	make	new	world	probably
9	person	make	way	don					day	think	world	good	world	think	know	think	know
10	day	life	say	day					life	actually	make	day	don	person	don	new	world
11	years	know	look	think					actually	know	new	bad	really	way	year	years	things
12	life	good	really	make					water	water	day	really	technically	water	think	day	think
13	called	having	know	water					say	years	person	person	water	know	way	human	earth
14	water	lot	world	actually					years	person	water	actually	human	actually	ve	know	going

covid

coronavirus

quarantine

virus

19

Evaluate

Top 15 Words Over Time

r/LifeProTips

	Oct '19	Nov '19	Dec '19	Jan '20	Feb '20	Mar '20	Apr '20	May '20	Jun '20	Jul '20	Aug '20	Sept '20	Oct '20	Nov '20	Dec '20	Jan '21	Feb '21
0	don	d8	don	don	don	time	don	don	don	don	don	don	don	don	don	don	don
1	use	d9	make	use	use	don	use	just	like	use	use	want	make	make	time	use	water
2	want	don	just	want	just	use	time	make	use	want	want	like	time	use	want	make	use
3	make	just	want	make	want	home	want	use	make	time	time	use	want	time	make	time	make
4	time	want	use	like	time	paper	like	want	time	make	make	make	use	just	just	just	just
5	phone	use	time	just	people	toilet	make	time	just	phone	like	need	just	people	use	people	want
6	just	phone	know	phone	phone	people	just	like	want	like	way	time	phone	good	need	want	time
7	d9	make	christmas	car	need	like	need	people	need	just	just	just	way	like	people	like	need
8	like	like	people	instead	make	make	try	instead	people	know	people	day	like	try	like	life	free
9	ll	time	instead	new	like	covid	phone	way	phone	need	know	people	need	want	ask	phone	people
10	d8	a7	like	time	way	quarantine	free	ll	instead	people	instead	new	people	need	way	ll	like
11	need	try	ll	try	stop	just	check	try	feel	ask	life	life	life	know	new	need	life
12	way	b1	save	water	try	19	save	know	ask	person	phone	phone	ll	ll	help	way	car
13	new	ll	new	check	day	need	car	ask	life	life	need	try	request	phone	know	know	look
14	car	need	water	ask	help	want	new	need	know	new	new	way	know	instead	instead	free	instead

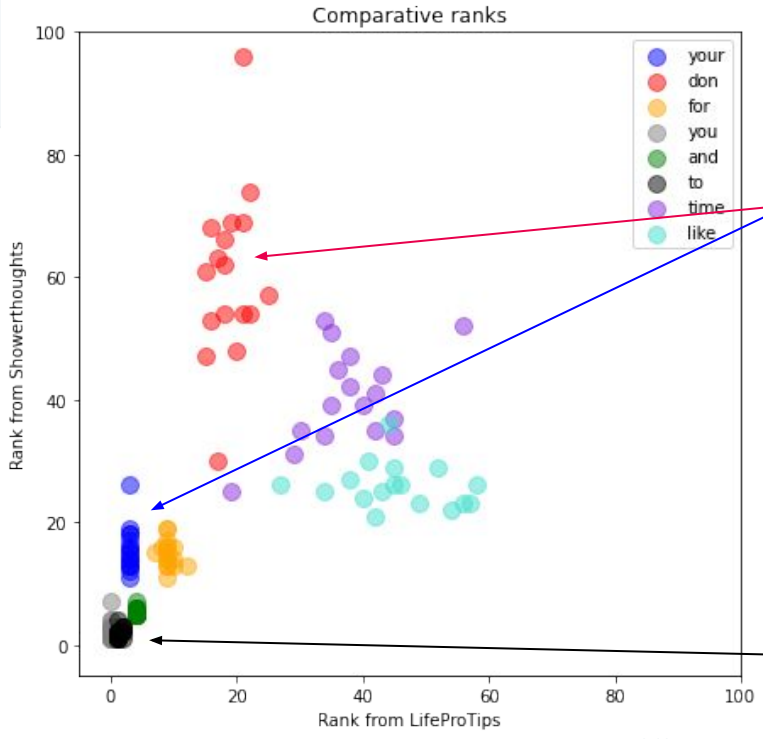
r/Showerthoughts

	Oct '19	Nov '19	Dec '19	Jan '20	Feb '20	Mar '20	Apr '20	May '20	Jun '20	Jul '20	Aug '20	Sept '20	Oct '20	Nov '20	Dec '20	Jan '21	Feb '21
0	people	people	people	just	people	people	people	people	just	people	just	people	people	people	people	people	people
1	just	just	just	people	just	just	just	just	people	just	people	just	just	just	just	just	just
2	like	like	like	like	like	time	like	like	like	like	like	like	like	like	like	like	like
3	probably	time	probably	play	time	probably	time	probably	time	probably	don	life	probably	time	time	life	time
4	time	probably	time	probably	life	world	probably	time	probably	life	time	probably	time	don	probably	time	don
5	don	don	good	life	don	like	don	make	really	time	probably	time	life	probably	life	don	life
6	world	think	life	time	probably	going	life	world	make	world	know	know	make	life	really	probably	make
7	know	day	person	man	day	covid	world	don	know	don	really	make	know	really	actually	make	mars
8	think	world	don	know	make	coronavirus	person	know	don	make	life	don	way	make	new	world	probably
9	person	make	way	don	know	right	say	person	day	think	world	good	world	think	know	think	know
10	day	life	say	day	think	quarantine	make	life	life	actually	make	day	don	person	don	new	world
11	years	know	look	think	really	virus	human	ve	actually	know	new	bad	really	way	year	years	things
12	life	good	really	make	say	19	think	right	water	water	day	really	technically	water	think	day	think
13	called	having	know	water	different	home	going	lot	say	years	person	person	water	know	way	human	earth
14	water	lot	world	actually	world	think	really	really	years	person	water	actually	human	actually	ve	know	going

Evaluate

Evaluate

Comparative Rank of Top 15 Words Over Time



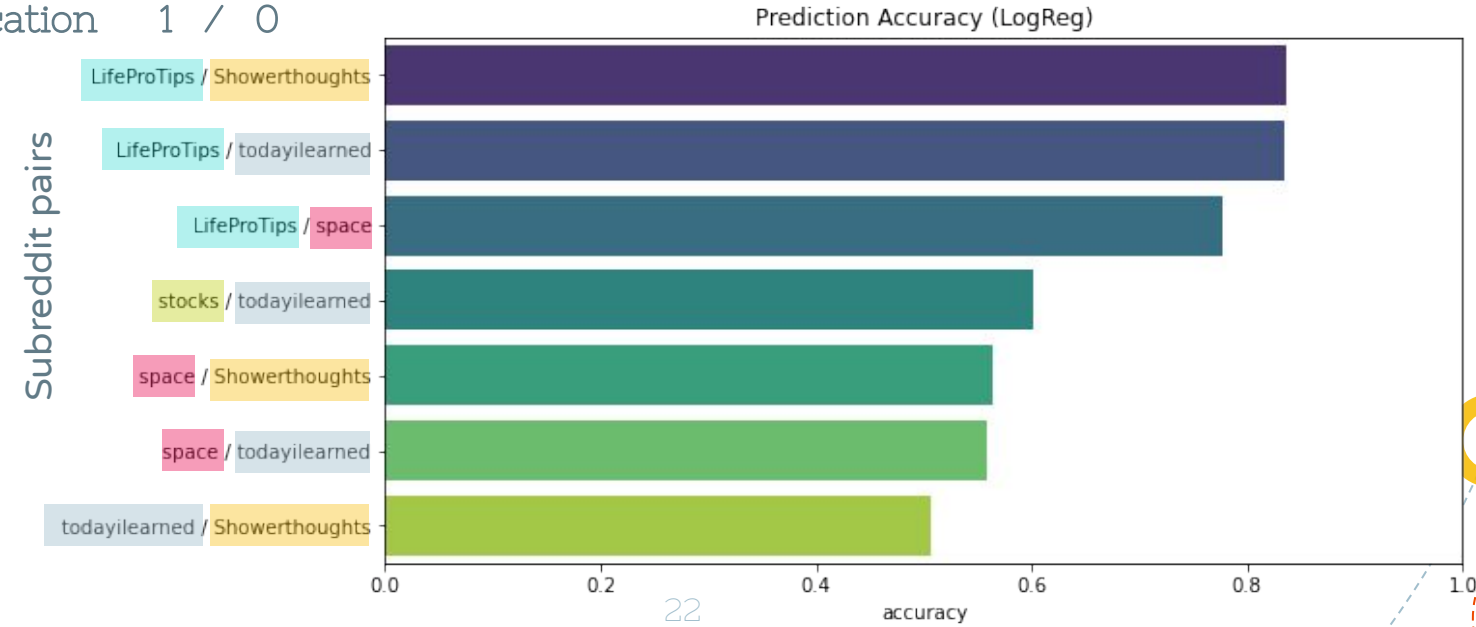
LogReg Coefficients

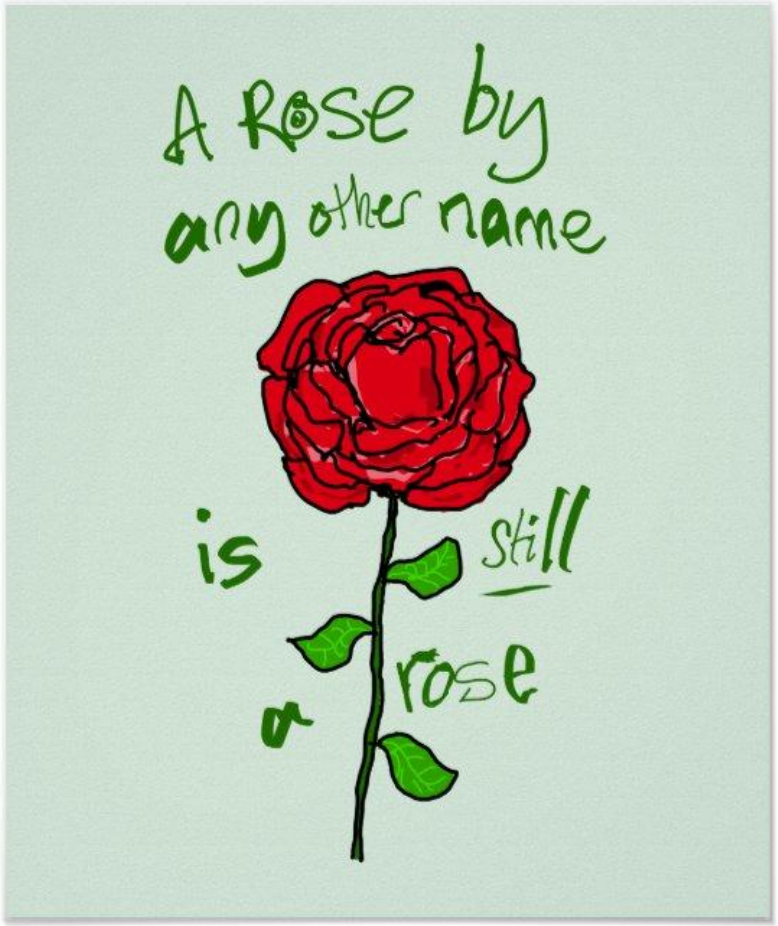
	Coefficient
your	126.027519
don	13.146453
use	9.150317
water	6.384216
for	6.292872
keep	6.159003
you	5.308579
want	5.239847
and	4.890368
to	4.889101

Is it possible to use the same model to differentiate between other subreddits?

More like what IS **r/LifeProTips**
and what IS NOT **r/LifeProTips**

Classification 1 / 0

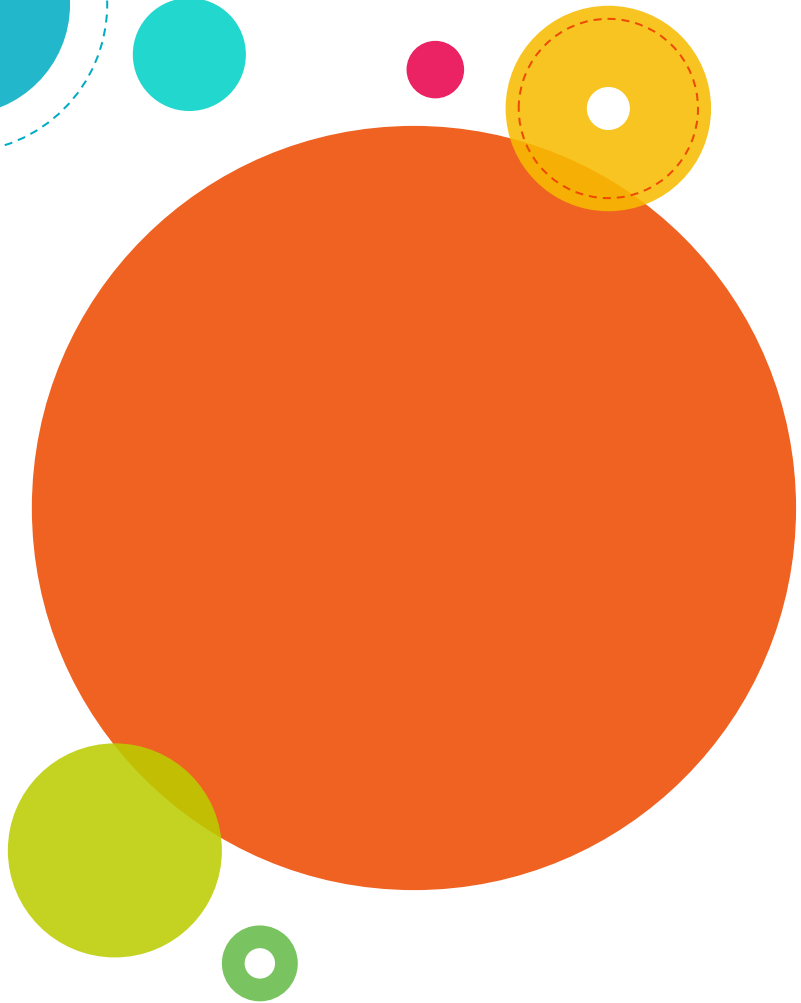



A Rose by
any other name
is still
a rose

Problem Statement: How well does a Classification model hold up over time?

Conclusions

- Models hold up quite well
- Seems applicable to the same subreddits over time
- My assumption is incorrect: if a word didn't appear in the training data, the model can't account for it
- Captures what IS and IS NOT **r/LifeProTips**
- **Future:**
  - Switch classification labels
  - Train model on 3 subreddits



QUESTIONS?

Aida Rahim
nuraida.rahim@gmail.com

Model

What Is The Effect of Not Removing Stopword?

74.5% of posts from the 'LifeProTips' contain the identifier 'LPT'