Plausible Utopia:
Scientists and Futurists on Reddit

Classification Modeling on Science and Futurology Subreddits

Allison Bishop

Background image source:
https://perthartglass.com.au/product/hs-mdi485sk-tremont-blue-streaky-opal-iridescent

What are we hoping to solve?

Who said it? A scientist or a futurist?

Both fields build and organize knowledge but the approaches are different

Does the vocabulary they use reflect those differences, enough to classify them?

Background image source:
https://www.canalplastic.com/products/mystique-blue-decorative-film-no-adhesive

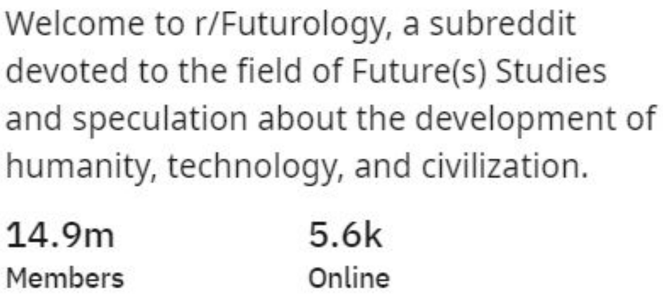
Methodology

1. Problem statement
2. Obtain data
3. Explore data
4. Model the data
5. Evaluate the model
6. Answer the original problem

The process is *not* as linear as it looks!

Background: The subreddits

r/futurology

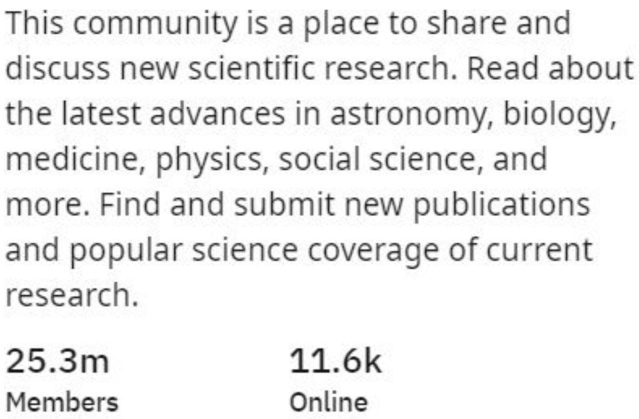


Welcome to r/Futurology, a subreddit devoted to the field of Future(s) Studies and speculation about the development of humanity, technology, and civilization.

14.9m
Members

5.6k
Online

r/science



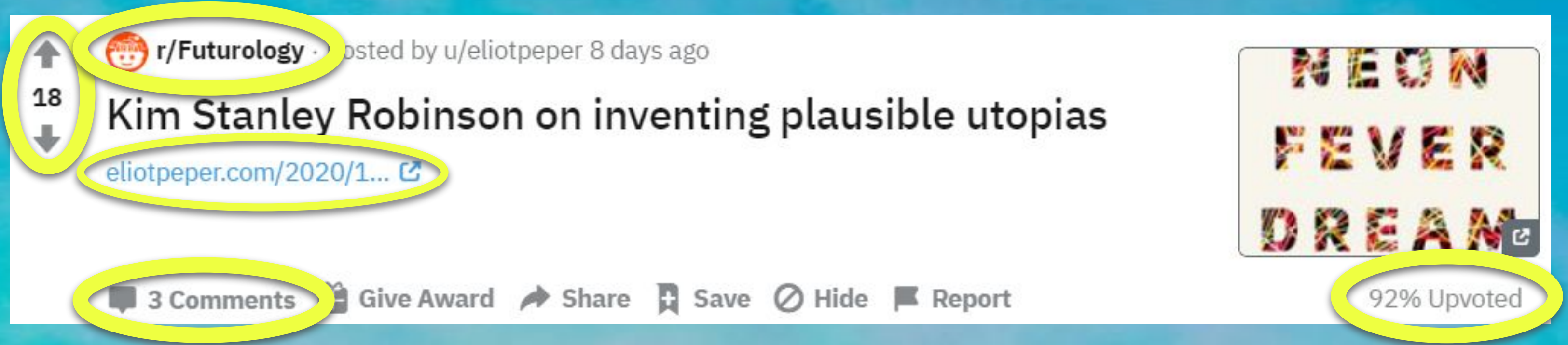
This community is a place to share and discuss new scientific research. Read about the latest advances in astronomy, biology, medicine, physics, social science, and more. Find and submit new publications and popular science coverage of current research.

25.3m
Members

11.6k
Online

*Data source: https://www.reddit.com/r/Futurology/, https://www.reddit.com/r/Science/

Anatomy of a subreddit post



r/Futurology · osted by u/eliotpeper 8 days ago

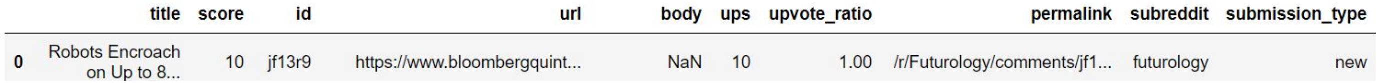
Kim Stanley Robinson on inventing plausible utopias

eliotpeper.com/2020/1...

Image source:
https://www.reddit.com/r/Futurology/comments/jagtts/kim_stanley_robinson_on_inventing_plausible/

About the data

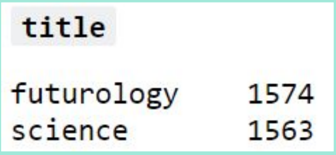
Raw data



	title	score	id	url	body	ups	upvote_ratio	permalink	subreddit	submission_type
0	Robots Encroach on Up to 8...	10	jf13r9	https://www.bloombergquint...	NaN	10	1.00	/r/Futurology/comments/jf1...	futurology	new

Pre-processing

Dimensions:
Rows: 3,137
Columns: 2

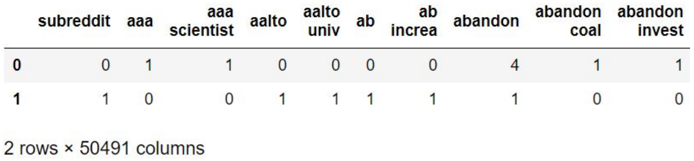


```
title

futurology    1574
science       1563
```

After processing

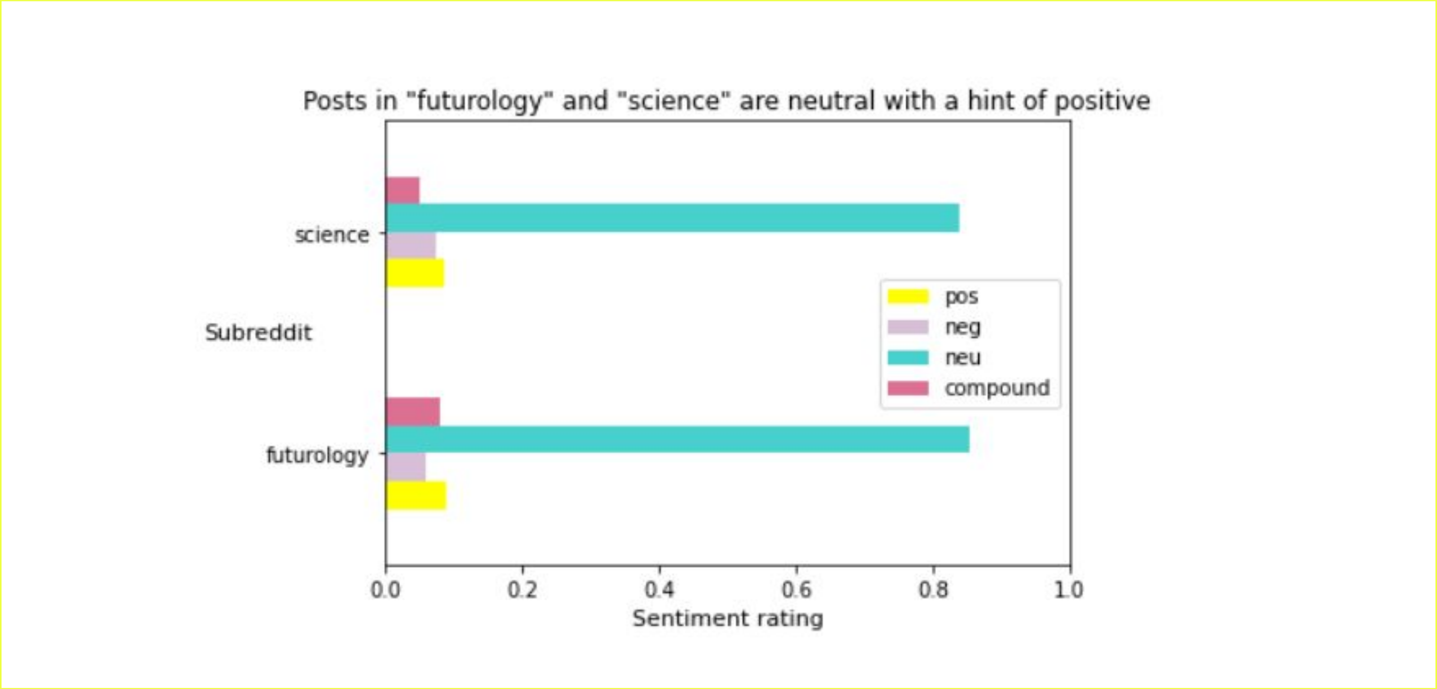
Dimensions:
Rows: 2
Columns: 50,491



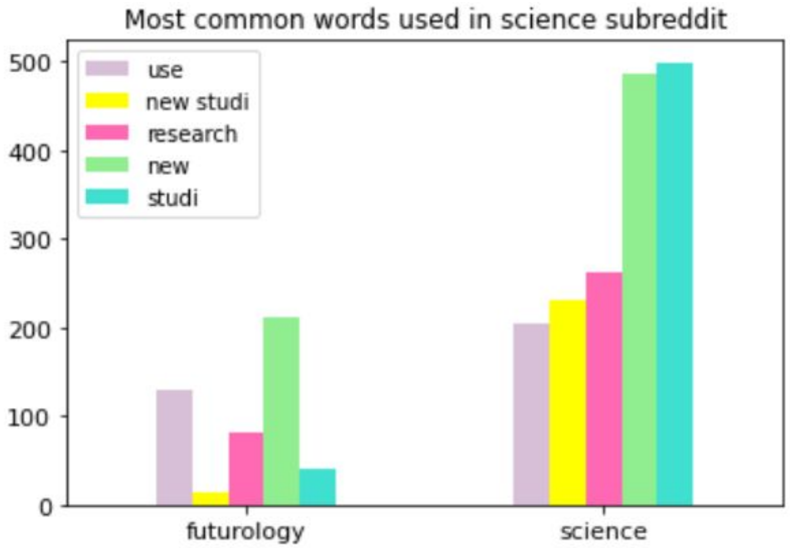
	subreddit	aaa	aaa scientist	aalto	aalto univ	ab	ab increa	abandon	abandon coal	abandon invest
0	0	1	1	0	0	0	0	4	1	1
1	1	0	0	1	1	1	1	1	0	0

2 rows × 50491 columns

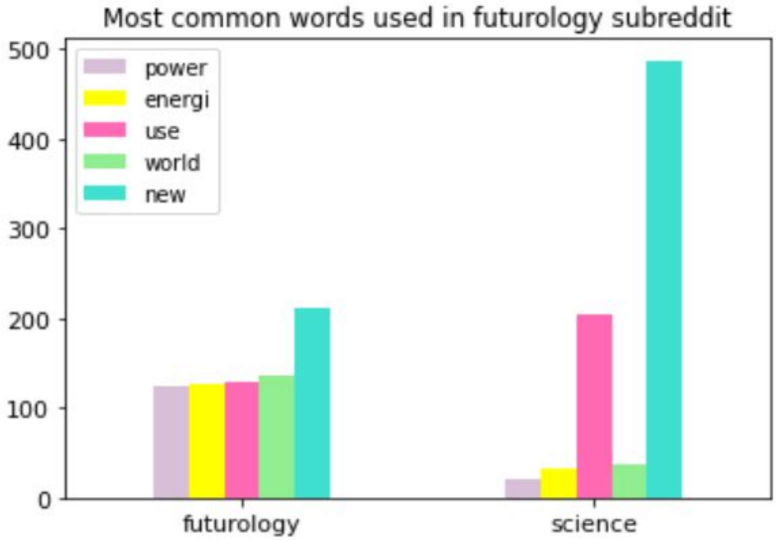
Data exploration



Data exploration



Data exploration



Models and measuring success

Supervised learning > Classification task > Logistic Regression and Random Forest models

Accuracy

How well the model makes predictions

Baseline accuracy score

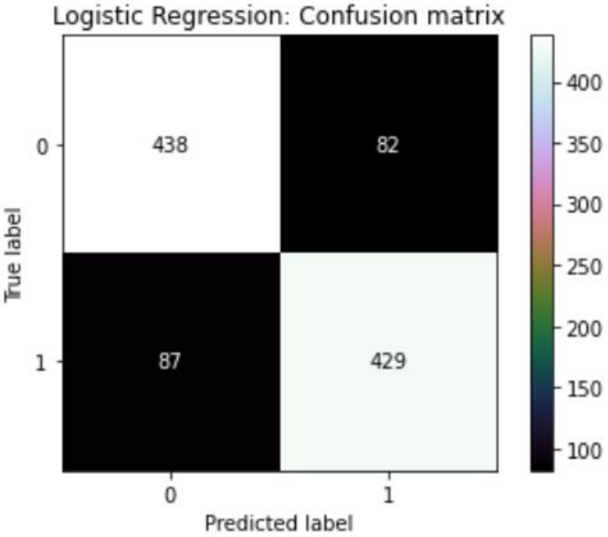
50.2%

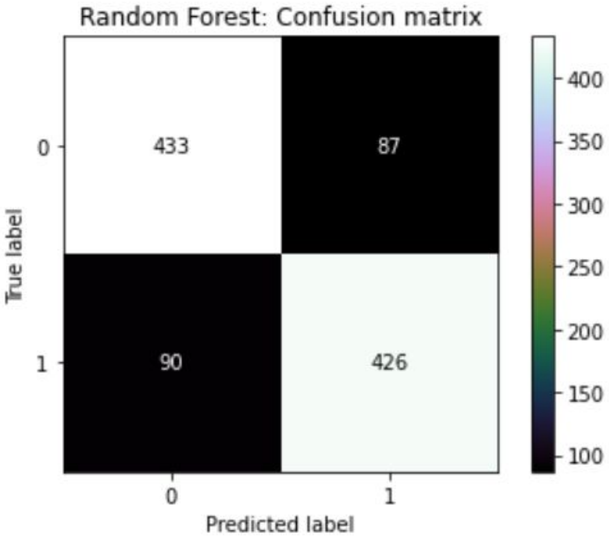
Comparison scores

Logistic Regression (ridge): 83.7%

Random Forest: 82.1%

Model performance





Logistic Regression
Accuracy: 83.7%

Random Forest
Accuracy: 82.1%

Next steps

Experiment with:

Feature engineering

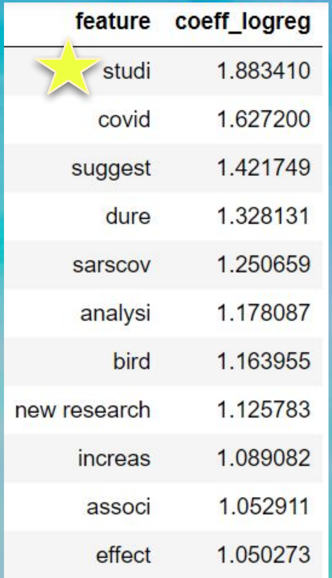
Boosting models

Topic modeling

Conclusions

The logistic regression model determined that the most important feature is ‘studi’

>> ‘studies’, ‘study’, ‘studied’, etc. <<



feature	coeff_logreg
studi	1.883410
covid	1.627200
suggest	1.421749
dure	1.328131
sarscov	1.250659
analysi	1.178087
bird	1.163955
new research	1.125783
increas	1.089082
associ	1.052911
effect	1.050273

* Holding all other variables constant