DISTINGUISHING BETWEEN SUBREDDITS

CHRIS SOKOLOWSKI

Problem Statement

How well can a data model distinguish a subreddit from a parody?
Will curating for popular posts lead to a more successful model?

r/TodayILearned

TIL that the F.B.I. and C.I.A. recruit heavily from the Mormon population because they are usually cheaper to do a security clearance on, they often speak another language from their mission trips and they usually have a low risk lifestyle.

Accurate and interesting facts

Title only, no body

Source required

Strict, sizable ruleset

No content on software/websites

Every post requires moderator's approval

r/ShittyTodayILearned

TIL The press used to mint coins is called a pounder and state quarters can have backs so low profile they require a special pounder fitted with a piece of cheese cloth to prevent the hot metal from sticking. This device is called a quarter pounder with cheese.

Jokes pretending to be accurate and interesting facts

Title only, no body

No sources

No reposts, no memes, no Rickrolling

A lot of posts mocking Reddit/internet culture

Almost zero moderation

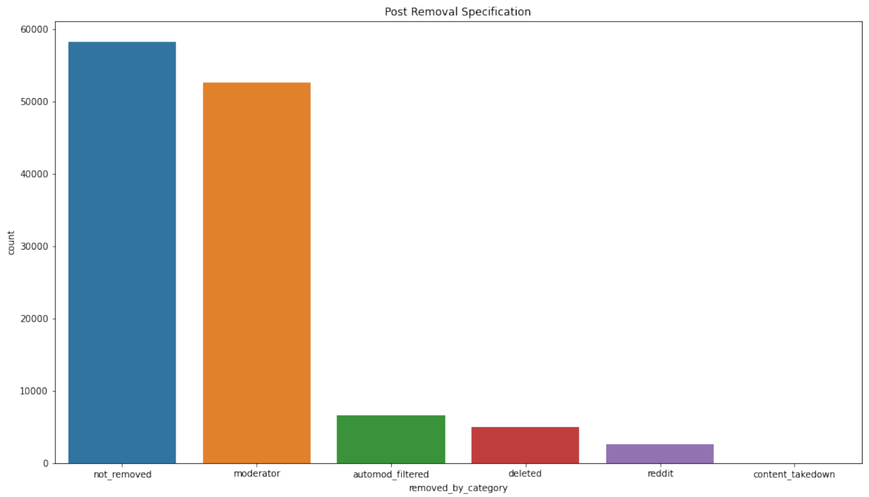
Method One

1. Scrape the entire r/STIL post history
2. Scrape an equal amount of posts from r/TIL
3. “Moderate” the r/STIL posts
4. Run multiple classification models
5. Create GridSearches for the most successful models

Method Two

1. Filter r/STIL posts to those with ten or greater upvotes
2. Grab an equal amount of r/TIL posts meeting the same requirement
3. Run multiple classification models
4. Create GridSearches for the most successful models

Remove the Removed Posts



53.4% of ~125,000 posts

No Posts about Reddit

TIL Reddit is a Genetic Wasteplant

TIFU by posting in the wrong subreddit

TIL Reddit isn't the only website on the Internet

TIL that the mods of r/shittylifeprotips think that boiling poptarts is a good idea.

TIL: Ive been taught more things in life that belong on this subreddit than belong on the actual TIL subreddit.

'reddit', 'subreddit', 'mod', 'moderator', 'username', 'r/', 'u/'

4.5% of ~7,800 posts

Words > 6

7: TIL that lava is manufactured in lavatories

6: TIL we didn't start the fire

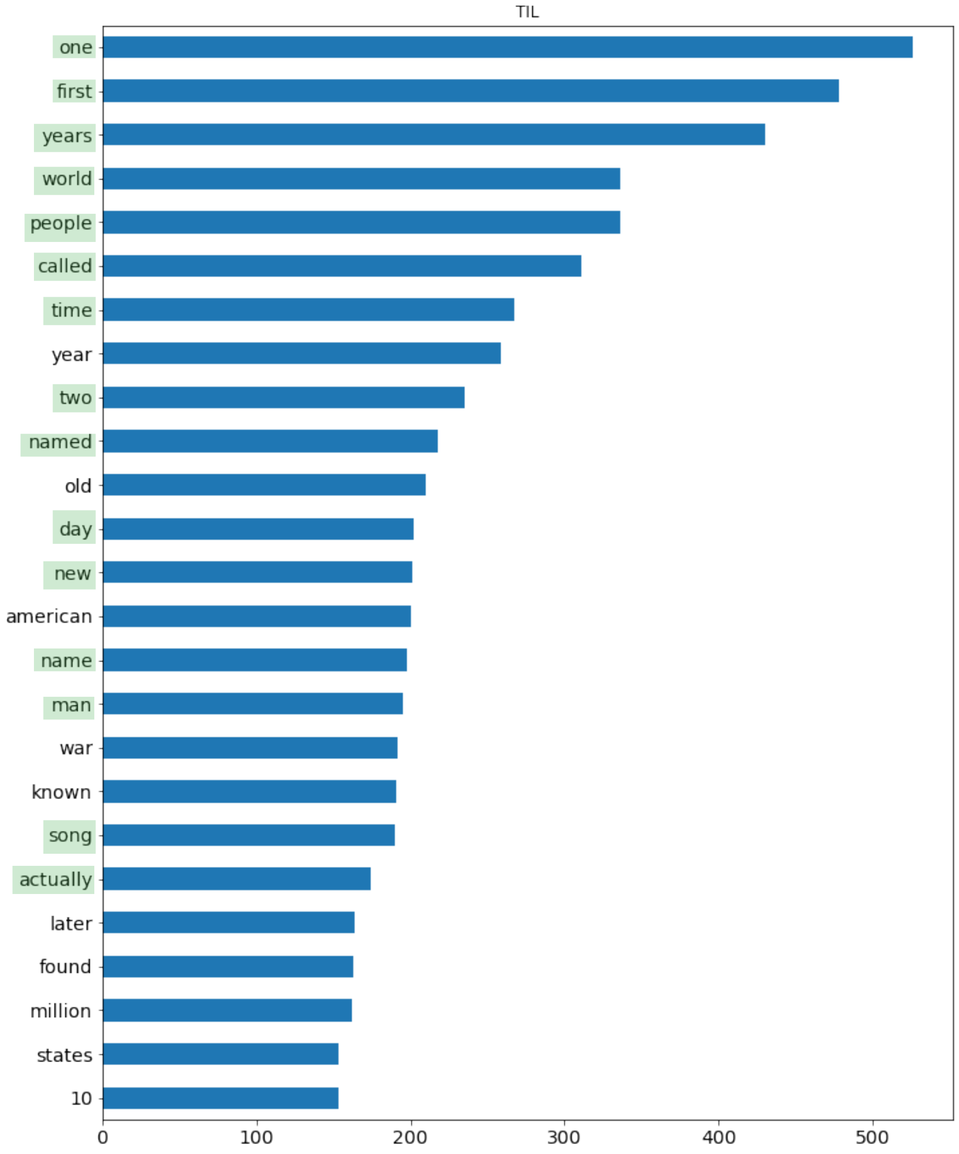
9.3% of ~7,500 posts

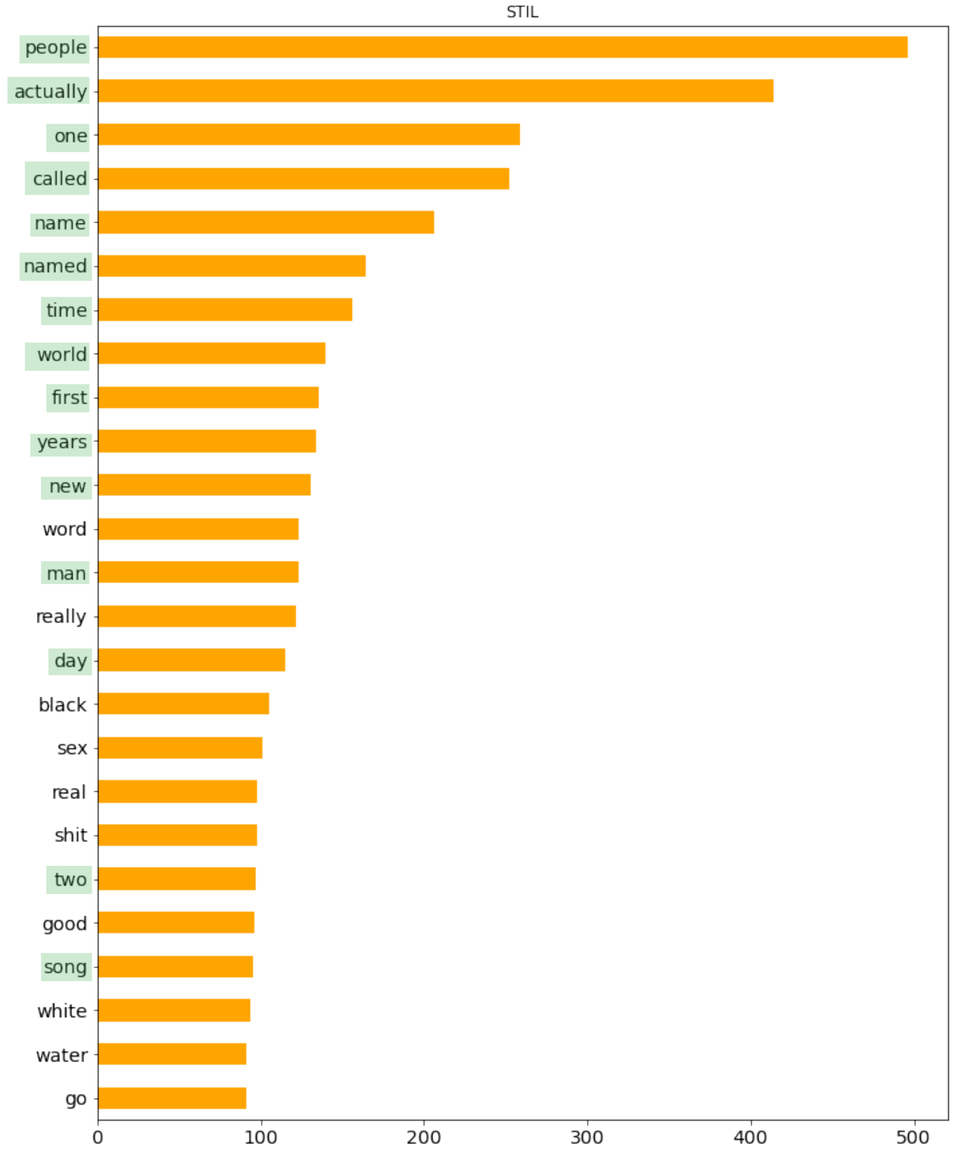
7: TIL lightning is hotter than the sun

6: TIL Alexander the Great smelled nice

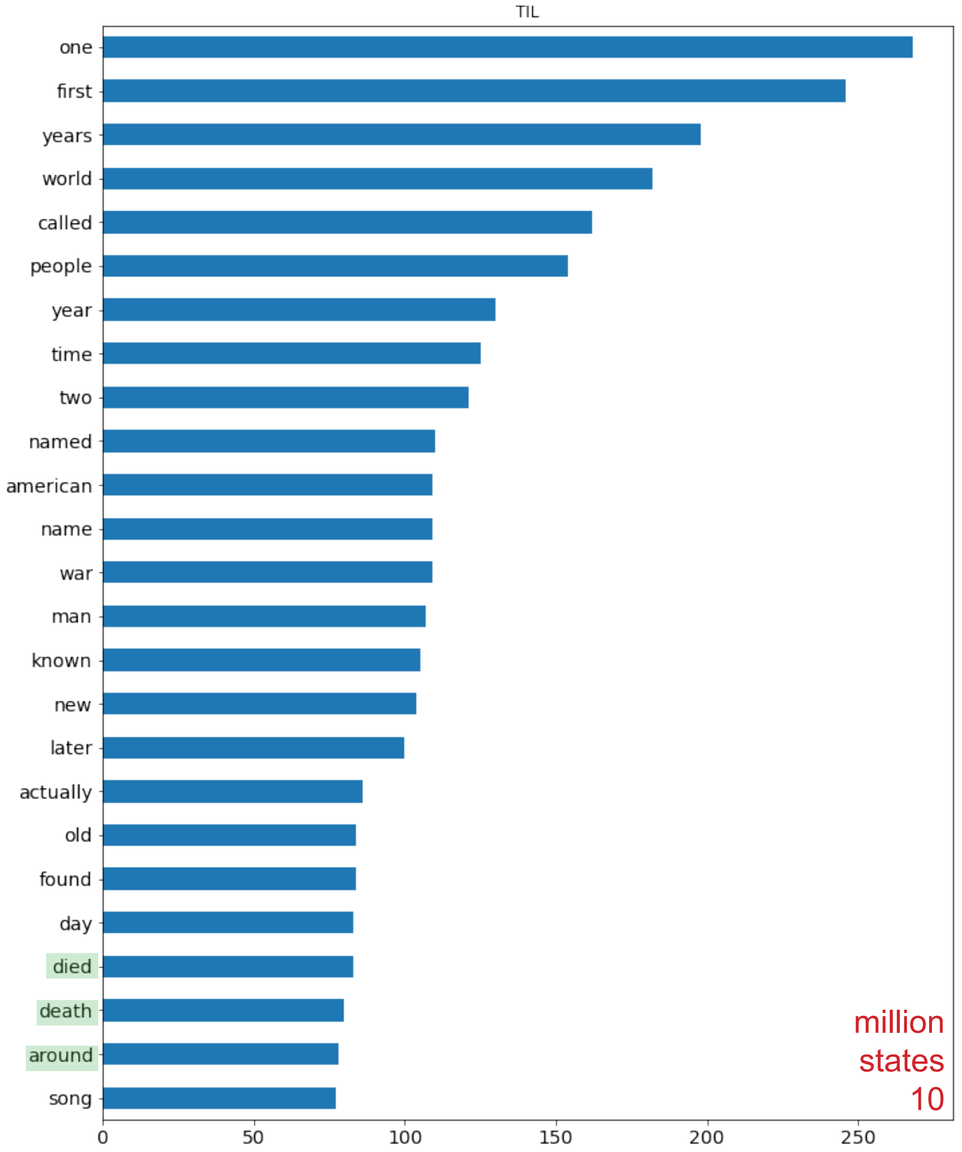
1.4% of ~58,000 posts

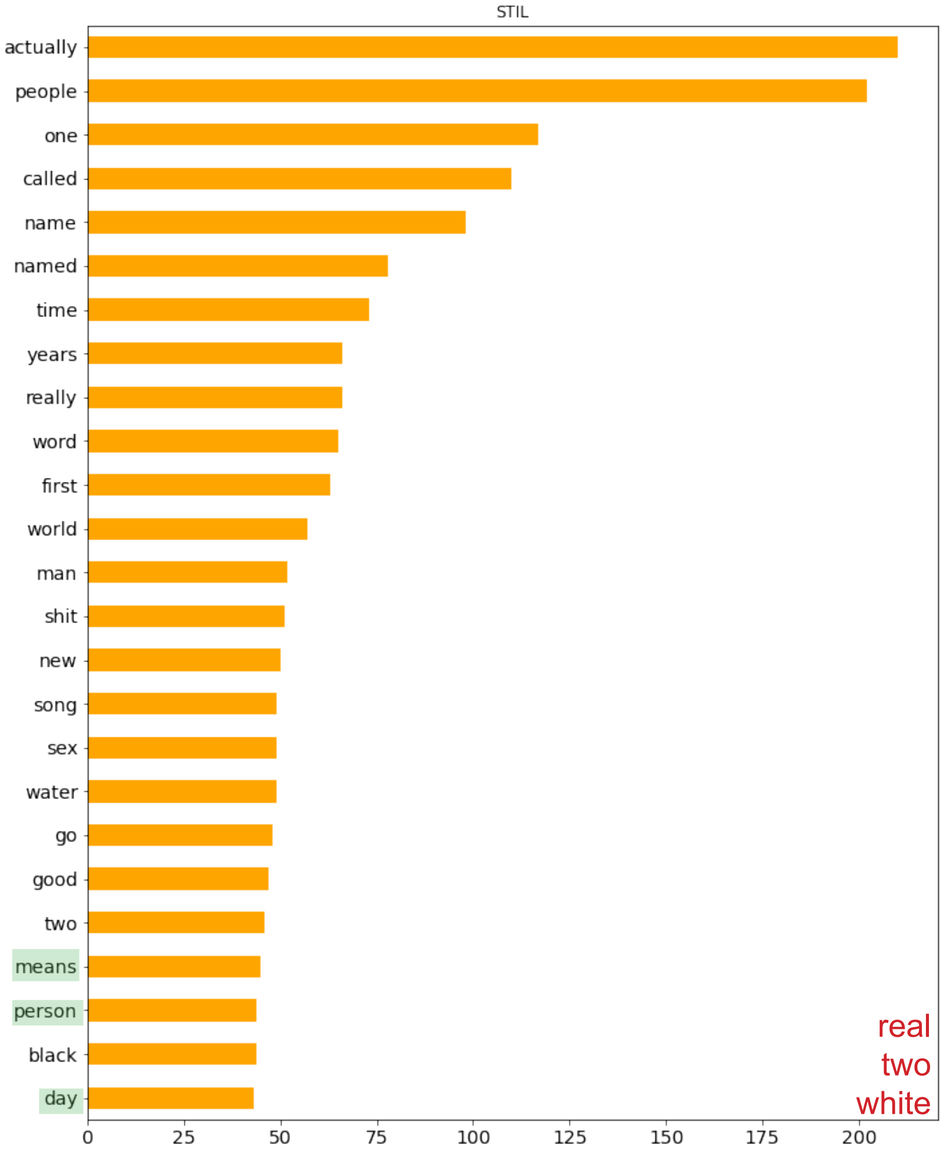
Method One: Top 25 Words



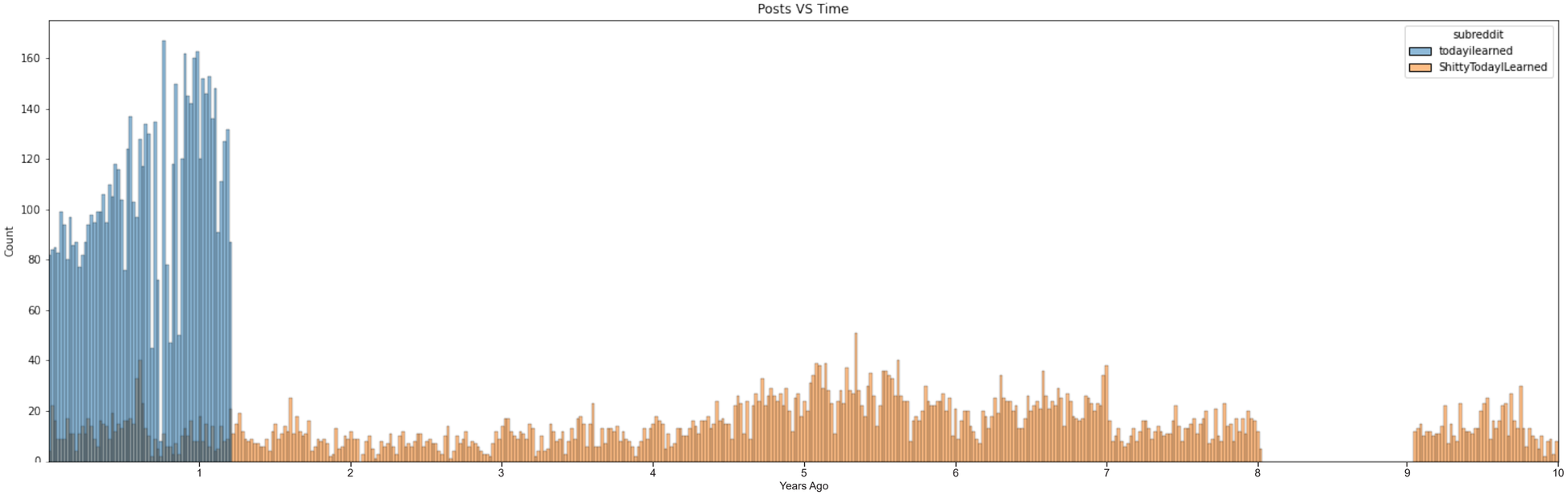


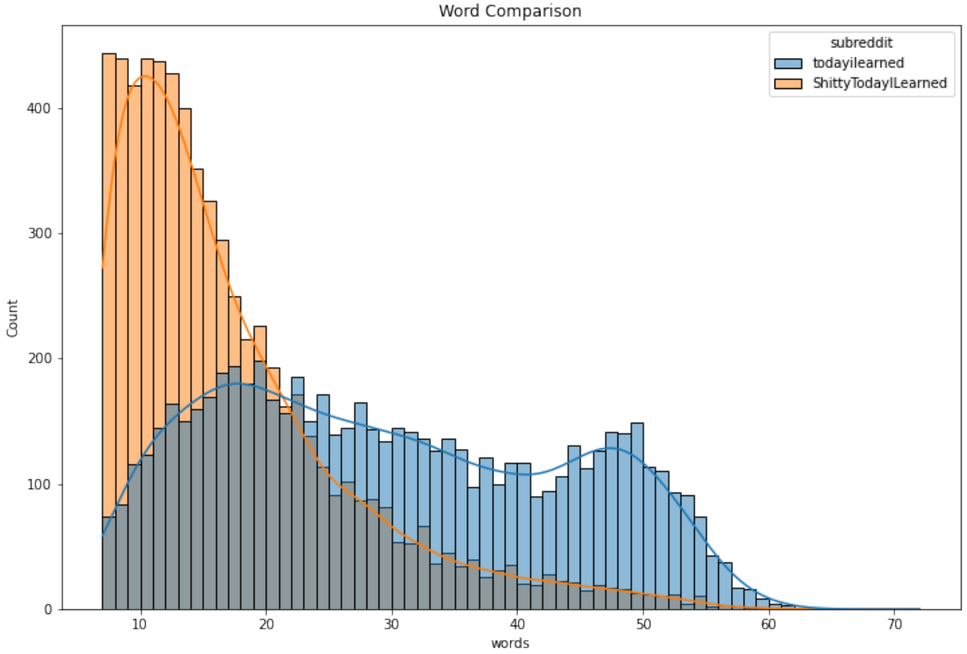
Method Two: Top 25 Words





Imbalances in Dataset





Method One

100% of 6,573 posts

11.7% of 56,149 posts

Method Two

45.0% of 6,573, 2,958 total

6.0% of 56,149 posts, 3,391 total (sampled 87.2%)

Method One

Classifier	Training Accuracy	Testing Accuracy
CV Logistic Regression	0.986	0.764
CV Decision Trees	0.999	0.672
CV Gradient Boost	0.716	0.671
CV ADA Boost	0.679	0.662
CV Bagging	0.975	0.694
TFID Logistic Regression	0.953	0.757
TFID Decision Trees	0.999	0.685
TFID Random Forest	0.999	0.683
TFID Extra Trees	0.999	0.753
TFID Gradient Boost	0.744	0.702
TFID ADA Boost	0.673	0.652
TFID Bagging	0.978	0.726

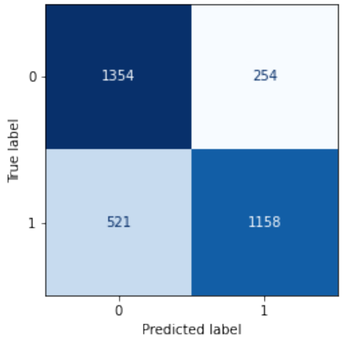
Method Two

Classifier	Training Accuracy	Testing Accuracy
CV Logistic Regression	1.000	0.766
CV Decision Trees	1.000	0.694
CV Gradient Boost	0.776	0.684
CV ADA Boost	0.730	0.674
CV Bagging	0.979	0.709
TFID Logistic Regression	0.981	0.761
TFID Decision Trees	1.000	0.705
TFID Random Forest	1.000	0.703
TFID Extra Trees	1.000	0.744
TFID Gradient Boost	0.792	0.708
TFID ADA Boost	0.726	0.661
TFID Bagging	0.983	0.733

Method One

Classifier	Training	Tr Diff	Testing	Te Diff
CV Logistic Regression	0.976	-1.02%	0.762	-0.26%
TFID Logistic Regression	0.887	-7.17%	0.749	-1.06%
TFID Extra Trees	0.993	-0.60%	0.760	0.92%

CV Logistic Regression

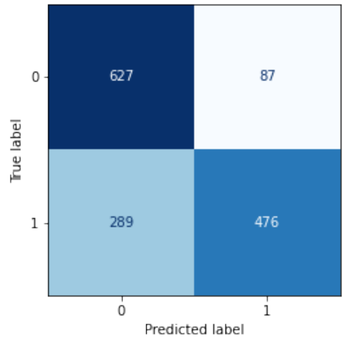


Specificity: 0.842
Precision: 0.820
Recall: 0.690

Method Two

Classifier	Training	Tr Diff	Testing	Te Diff
CV Logistic Regression	0.985	-1.51%	0.771	0.65%
TFID Logistic Regression	0.919	-6.53%	0.763	0.26%
TFID Extra Trees	0.872	-13.68%	0.773	3.82%

TFID Extra Trees



Specificity: 0.877
Precision: 0.845
Recall: 0.629

In Conclusion

Method 2 (TFID Extra Trees, 77.3%) slightly outperformed Method 1 (CV Logistic Regression, 76.0%)

Almost all classifiers were extremely overfit

Most classifiers fell around 70% prediction rate, compared to 50% baseline

Must be a lot of overlap between subreddits

Model lacking ability to “fact check”

Get a dataset of r/TIL over the past 10 years

Method 2 could possibly be better optimized

Beat The Model

TIL of Yoshie Shiratori, who escaped from prison 4 times. First, he picked a lock, then he climbed the walls and escaped through a skylight. Third, through a narrow food slot in his door, and finally he dug a tunnel out. He was recaptured each time, and eventually released for good behavior.

TIL after an obese umpire died during a game, Major League Baseball decided to enforce weight limits. In 1999 under this policy, umpire Eric Gregg was fined $5,000 for exceeding 300lbs.

TIL The original American Transcontinental Railroad was a major undertaking which took many years to build. It was built for the sole purpose of making it easier to move people and stuff.

TIL about Null Island, the place where the Prime Meridian and Equator meet. It is a fictitious island with the geographical coordinates of 0,0 - off the coast of Africa. And it only exists in geographical databases to highlight data errors.

Beat The Model

TIL of Yoshie Shiratori, who escaped from prison 4 times. First, he picked a lock, then he climbed the walls and escaped through a skylight. Third, through a narrow food slot in his door, and finally he dug a tunnel out. He was recaptured each time, and eventually released for good behavior.

TIL after an obese umpire died during a game, Major League Baseball decided to enforce weight limits. In 1999 under this policy, umpire Eric Gregg was fined $5,000 for exceeding 300lbs.

TIL The original American Transcontinental Railroad was a major undertaking which took many years to build. It was built for the sole purpose of making it easier to move people and stuff.

TIL about Null Island, the place where the Prime Meridian and Equator meet. It is a fictitious island with the geographical coordinates of 0,0 - off the coast of Africa. And it only exists in geographical databases to highlight data errors.