NLP

An analysis of predictability using word-text and Machine Learning models.

Problem Statement:

How accurately can the words in a subreddit's post predict whether the post was intended for a 'good advice' or 'bad advice' subreddit?

Subreddits

LifeProTips

"Tips that improve your life in one way or another"

A subreddit dedicated to sharing 'helpful' user-provided advice for navigating a number of different kinds of situations. These are typically akin to 'Life Hacks'.

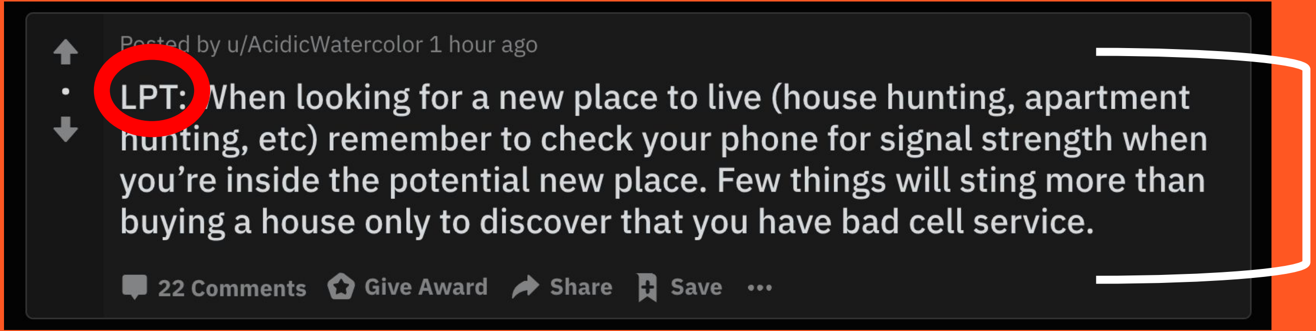
17+ Million Subscribers

UnethicalLifeProTips

"An Unethical Life Pro Tip (or ULPT) is a tip that improves your life in a meaningful way, perhaps at the expense of others and/or with questionable legality. *Due to their nature, do not actually follow any of these tips–they're just for fun. Share your best tips you've picked up throughout your life, and learn from others!"*

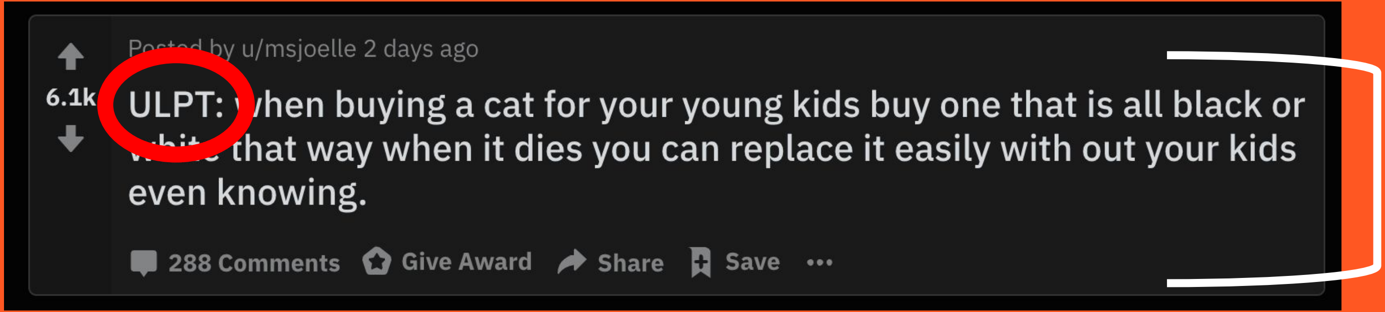
1+ MIllion Subscribers

The Raw Data:
Background



Posted by u/AcidicWatercolor 1 hour ago

LPT: When looking for a new place to live (house hunting, apartment hunting, etc) remember to check your phone for signal strength when you’re inside the potential new place. Few things will sting more than buying a house only to discover that you have bad cell service.



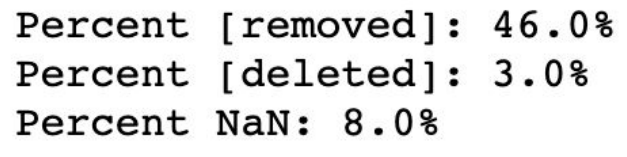
Posted by u/msjoelle 2 days ago

ULPT: when buying a cat for your young kids buy one that is all black or white that way when it dies you can replace it easily with out your kids even knowing.

The Data: Prepping the models

Pre-Processing with stopwords, lemmatizing, and removing non-letters.

- **Title = Key Feature for modeling.**
  - On just one 1 API pull, the Selftext field had over 50% missing or ‘odd’ values



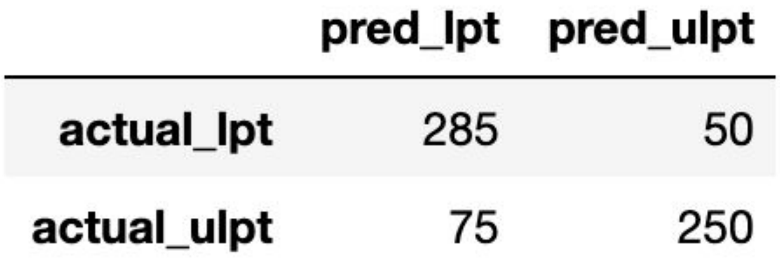
```
Percent [removed]: 46.0%
Percent [deleted]: 3.0%
Percent NaN: 8.0%
```

- **Some Words Removed:**
  - lpt
  - ulpt
  - lptrequest
  - ulptrequest

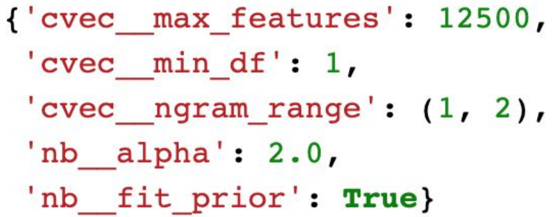
To the Notebook.....

The Multinomial Naive Bayes scored the highest accuracy across all test.

All of the models were overfit though.



	pred_lpt	pred_ulpt
actual_lpt	285	50
actual_ulpt	75	250



```
{'cvec__max_features': 12500,
 'cvec__min_df': 1,
 'cvec__ngram_range': (1, 2),
 'nb__alpha': 2.0,
 'nb__fit_prior': True}
```

0.97 on Train
0.79 on Test

0.80 Best Score