Is your advice unethical?

Scott Atkinson

Introduction

How do we detect unethical advice?

General themes indicated by data

LifeProTips: self-improvement

> *"hold both ends of the tube and run it over the edge of the sink to push toothpaste to the top - you'll get almost every last bit with almost no effort"*

UnethicalLifeProTips: dishonesty with the purpose to gain from others

> *"The best way to hang up on someone is in the middle of your own sentence. That way they never suspect you of hanging up on them."*

Uses

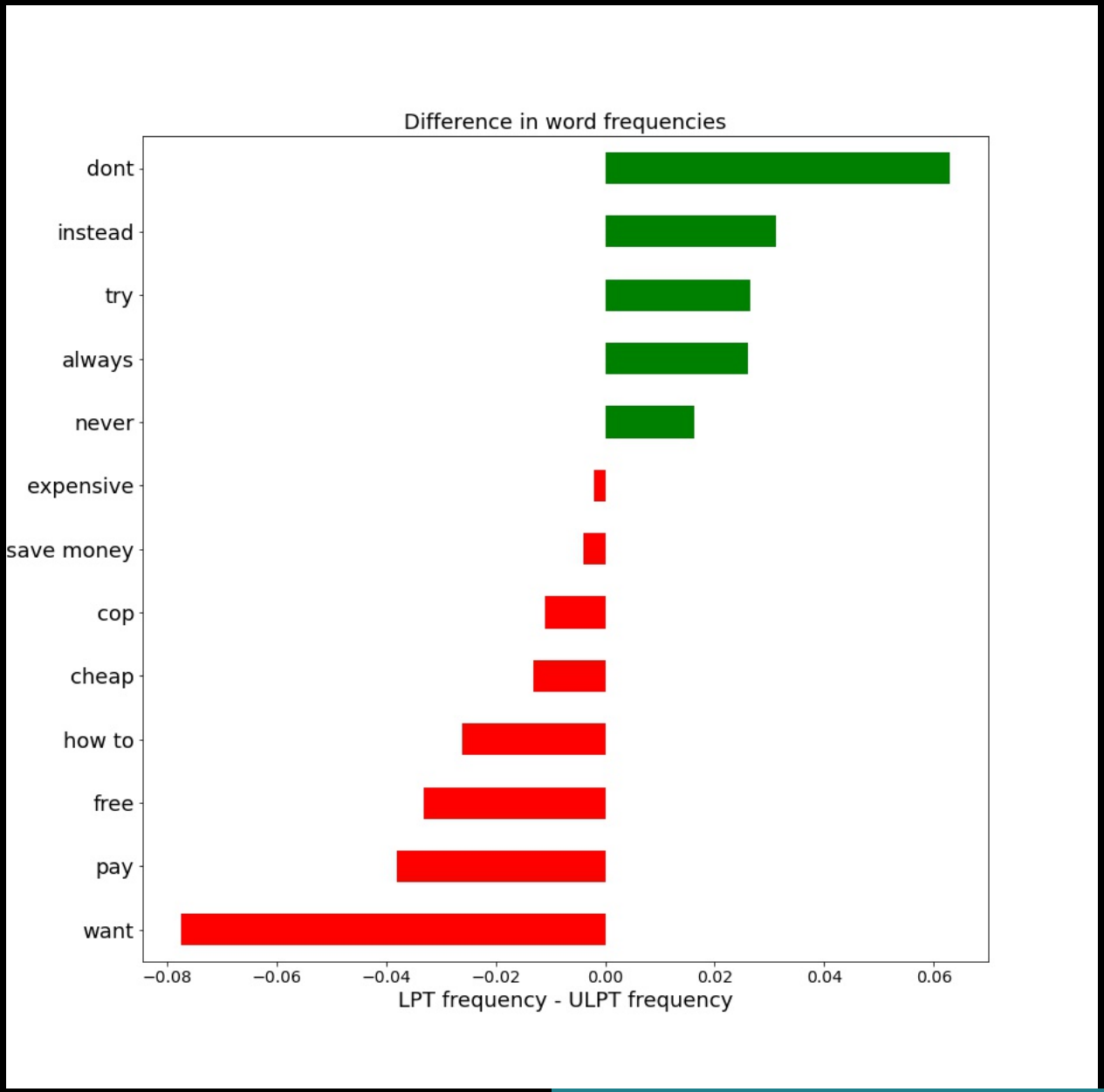
- automate moderation of social media/online forums
- automate monitoring of internal business communications
- tool for public relations

Data sources

LifeProTips and UnethicalLifeProTips subreddits:

- `subreddit`
- `title`
- `id`
- `created_utc`
- `score`
- `num_comments`
- `selftext`

EDA: expected words

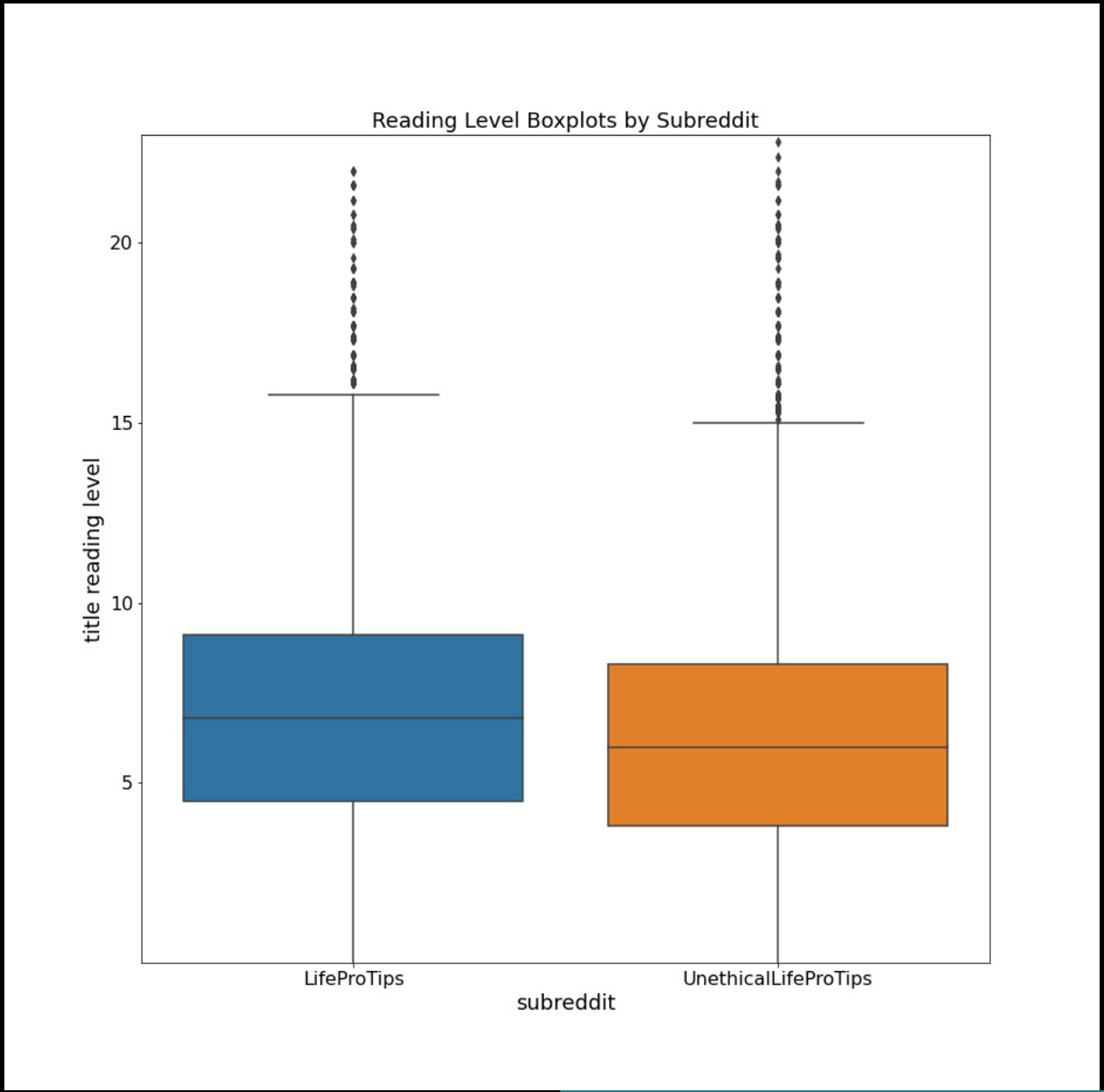


EDA: reading level

Table: Flesch-Kincaid grade levels for `title`

subreddit	mean	std	Q1	median	Q3
LifeProTips	7.0028	3.5783	4.5	6.8	9.1
UnethicalLifeProTips	6.3536	3.8397	3.8	6.0	8.3

EDA: reading level



EDA: Clustering LPT topics

All LPT topics align with self-improvement theme.

EDA: Clustering LPT topics

Resolutions: The data were scraped from Reddit in early January, so there were understandably a large number of recent posts addressing New Year's resolutions. Here are some samples from this topic:

> *"Don't wait for the first day of the new year to start or change something. TODAY is the day to start, no matter where it falls on the calendar!"*

Cooking/kitchen tips: There was significant representation of tips and advice regarding cooking and preparing food. Some notable samples are:

> *"Don't use the microwave to heat up pasta from the fridge, but use a frying pan and a little bit of water"*

EDA: Clustering LPT topics

Money: Tips about money management were often clustered together. Samples:

> *"if you're about to spend a good chunk of money on something you don't really need, stop and invest the money instead."*

Gift giving: Again, because the data were scraped near the holidays, there were many posts referring to Christmas gifts. Some samples include:

> *"Keep an ongoing list of potential Christmas presents for your loved ones throughout the whole year. Every time they mention something they'd like to have, write it down"*

EDA: Clustering LPT topics

Mental health: This is not a surprising topic to have received attention in the LPT subreddit. Some examples of posts are the following:

> *"It's okay (and often healthy) to take a break from listening to drama and overhyped, overinflated news/opinions on social media and mainstream news stations"*

Cleaning: Many tips surrounded the topic of cleaning. Here are a few examples:

> *"When cleaning the house, if you're bringing something from one room to another, bring back something that belongs to the room you were in."*

EDA: Clustering LPT topics

Online behavior: Another common theme is advice regarding online behavior: managing online profiles, accounts, passwords, etc. Samples:

> *"have 3 passwords: one for your main mail account, one for websites with your card info, one for the other"*

EDA: Clustering ULPT topics

All ULPT topics align with the dishonesty-for-gain-from others theme.

EDA: Clustering ULPT topics

Avoid ads/paywalls: This topic includes tips on how to avoid ads or paywalls on various websites. Samples:

> *"Press Control+A, then Control+C to highlight and copy an article before the paywall pops up. Then paste it into a Google or Word Document to read it."*

Scamming return policies/rewards programs: Another large class of ULPT posts included advice on how to exploit the return policies and rewards programs of various businesses. Examples include:

> *"If you need to get a new refrigerator filter, buy a new filter, then put the old filter back in the package and return it saying you bought the wrong one."*

EDA: Clustering ULPT topics

Get out of work: There is a lot of advice in ULPT on how to get away with doing little to no work at your job. Examples:

> *"Working from home and need to appear online? Prop up a lock on the period button within the note pad application."*

Interpersonal deception/spite/prank: This is a broad category. Behavior including deception, eavesdropping, lying, manipulation, emotional abuse is covered by this category. The tips in this category are less for monetary or material gain and more for spite or some sort of social/emotional/intangible advantage over another individual. Some examples include:

> *"When you have guests over make the medicine cabinet full where if they try to open it things will fall out and you'll hear the noise. Then you will know if they're snooping around"*

EDA: Clustering ULPT topics

Getting something for nothing: This topic includes advice for how to get something for free or for less than full price/effort. Samples:

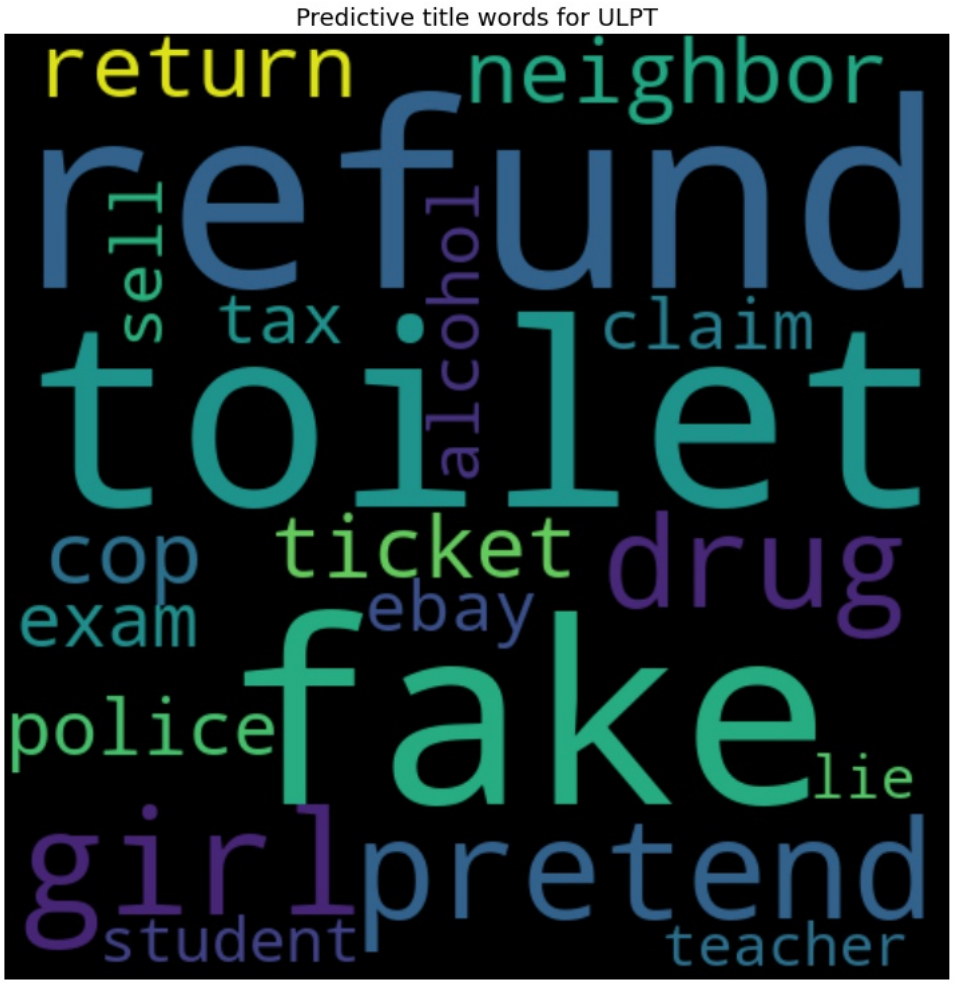
> *"If shipping packages through USPS, use the self service checkout and when you weigh your item, lift the corner of your package off the scale for a cheaper shipping rate"*

Car related: There is also a large representation of advice regarding getting out of traffic/parking violations and similar situations. Examples:

> *"Pass a cop while speeding on the highway and they start turning around to pull you over? Call 911 and report a drunk driver a few miles behind you."*

EDA: Most predictive title words for LPT

EDA: Most predictive title words for ULPT



Model selection

Table: Cross-validation results

Estimator	Hyperparameters	Mean CV Accuracy	Test Accuracy
`RandomForestClassifier`	`criterion='gini', max_depth=78, max_features='log2', n_estimators=143`	0.7614	0.7880
`MultinomialNB`	`alpha=1`	0.7772	0.7917
`LogisticRegression`	`C=0.1, dual=False, penalty='l2', tol=0.0001`	0.7714	0.8022

Threshold tuning

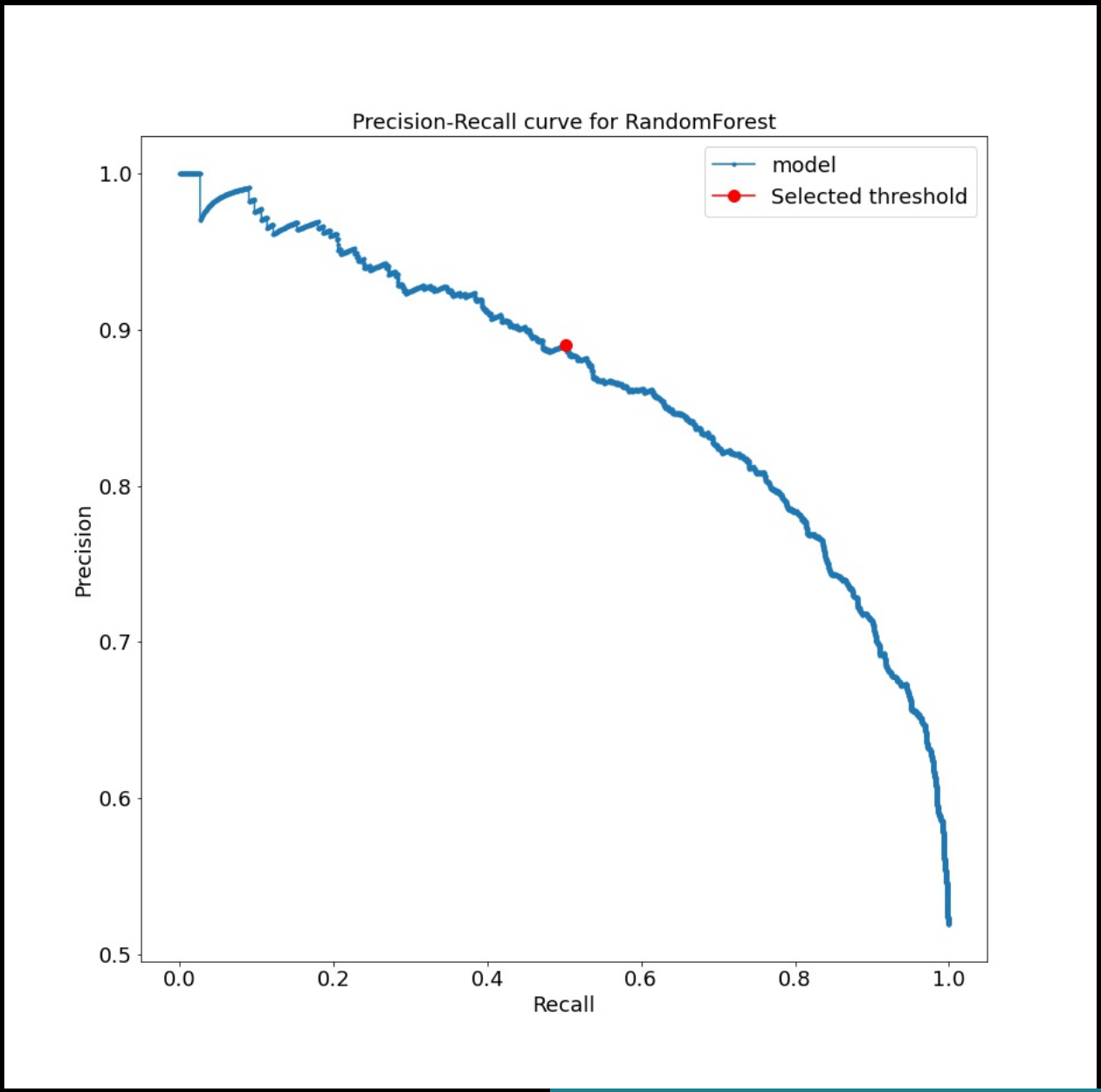
Unethical advice filter: emphasize precision.

Selected model: `RandomForestClassifier` with threshold 0.5718.

Positive precision: 89.03%

Positive recall: 50.2%

Precision-recall curve



Confusion matrix

Table: Confusion matrix for `RandomForestClassifier` with threshold 0.5718 predicting ULPT

	Predicted False	Predicted True
Actual False	47.0%	3.2%
Actual True	24.6%	25.2%

Summary

- Use: automatically detect unethical advice
- Precision: 89.03%
- Recall: 50.20%

Next steps

- Use more data
- Automate collection and training