Know your shit!

Life Pro Tips vs Shitty Life Pro Tips

By: Paul Cyr

Goal and Data Collection

Goal: To create model to predict which subreddit a post came from, Life Pro Tips or Shitty Life Pro Tips

Data Collection: Web scrap subreddits r/LifeProTips and r/ShittyLifeProTips to get 500 posts of each subreddit

Life Pro Tip or Shitty Life Pro Tip
Can you tell the difference?

-If you're at a wedding or some other event, set your Tinder search radius to 1 km to see who's single
OR
-Swallow magnets to become attractive

-Download your account data from Google to see in reality what they store on you
OR
-If you're worried about your Facebook Data being misused, just switch to Instagram

Data Cleaning

Stop_words:

- English stop_words
- Subreddit specific words like LPT, SLPT
- Number ‘words’
- Word that were common in both dataframes: 'time', 'like', 'make', 'want', 'use’ (*not very effective)

Top 10 words by subreddit (mean)

LPT

target_	0
amp	0.056
people	0.048
know	0.048
way	0.046
tell	0.046
car	0.044
water	0.038
work	0.038
need	0.034
money	0.034

SLPT

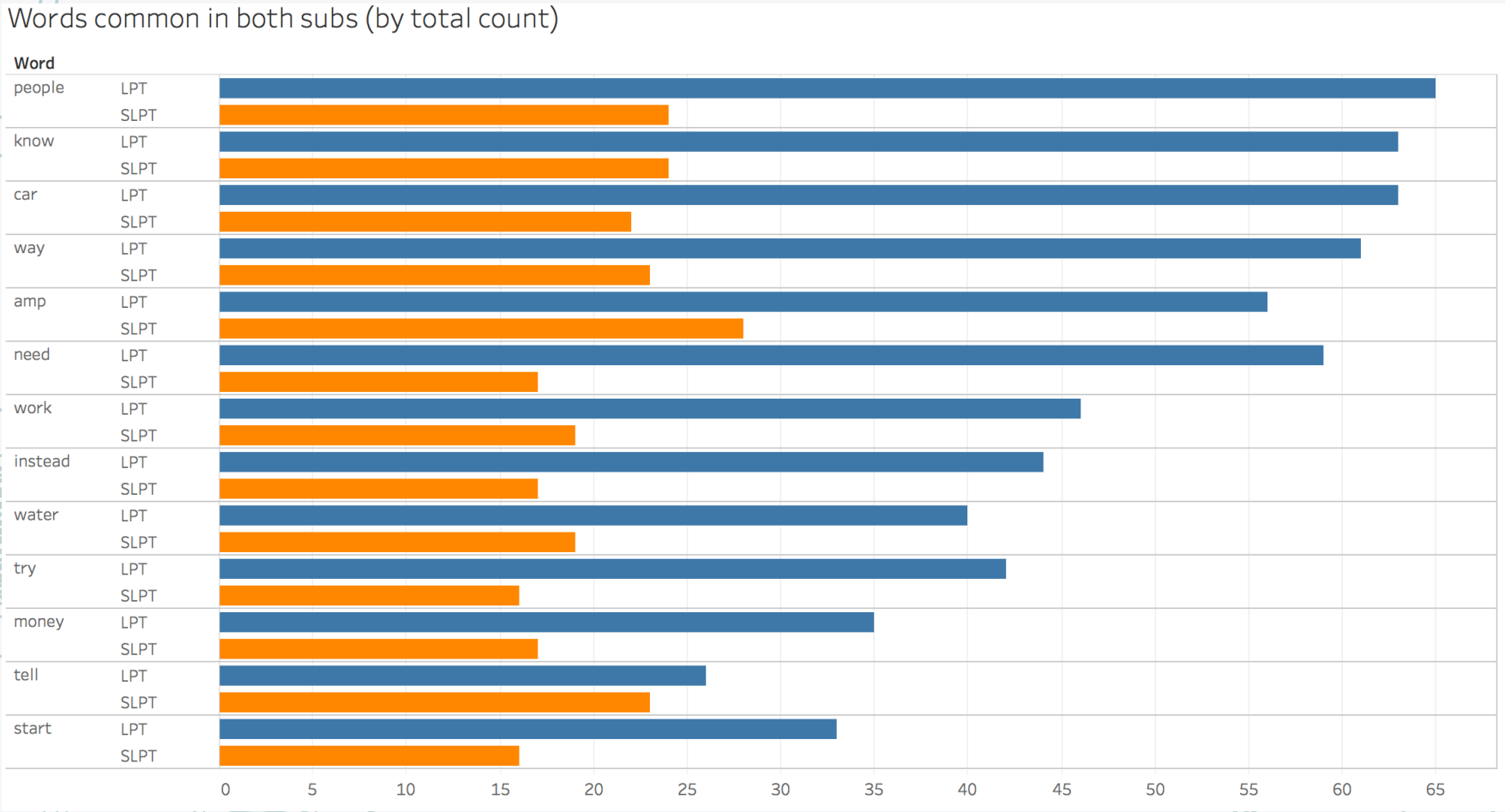
target_	0
people	0.048
know	0.048
car	0.044
way	0.046
need	0.034
amp	0.056
ask	0.016
say	0.020
day	0.024
work	0.038

Words that are in top 10 of both subreddits:

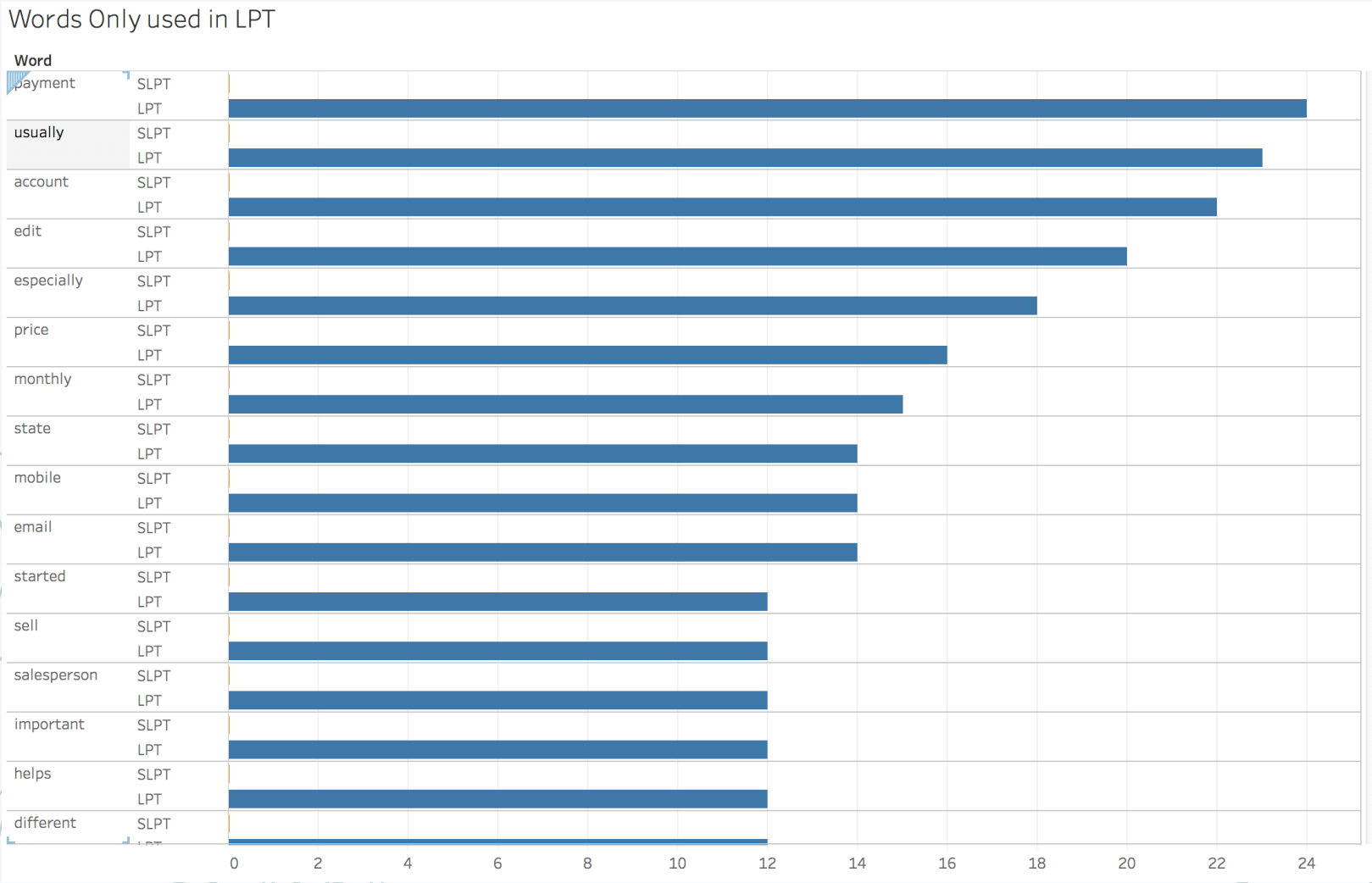
- ‘people’
- ‘know’
- ‘way’
- ‘car’
- ‘work’
- ‘need’

Removing common words in both subreddits was not effective at increasing score

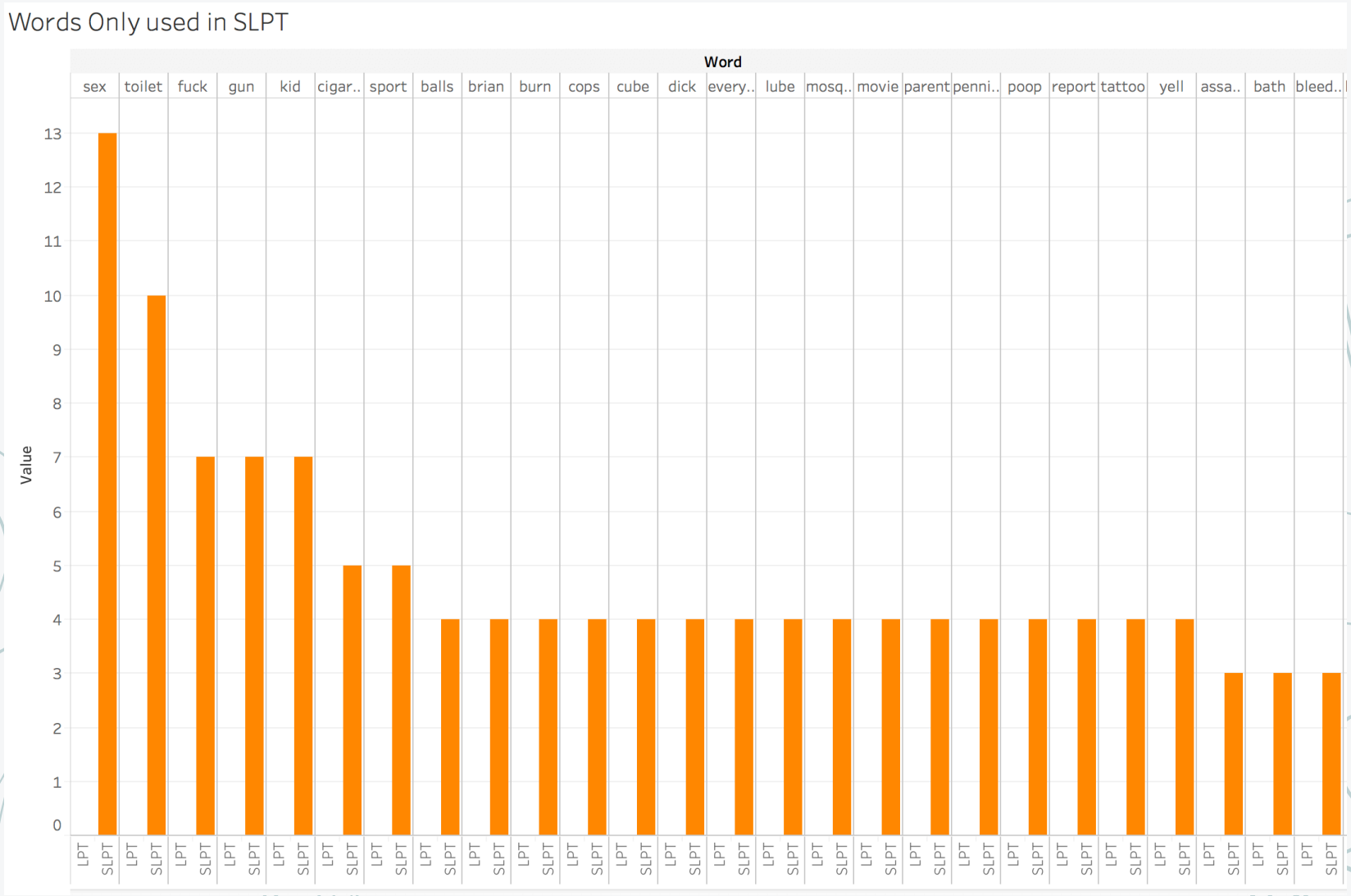
Top words by subreddit (total count)



Words only used in LPT



Words only used in SLPT



Data Modeling

Models and Score

Models	y	score
LogisticRegression using CountVectorizer (non-binary)	(title)	0.728
LogisticRegression using CountVectorizer (binary)	(title)	0.734
LogisticRegression using TfidfVectorizer	(title)	0.833
RandomForest using TfidfVectorizer	(title)	0.767
Pipeline StandardScaler and LogisticRegression	(ups and num_comments)	0.796

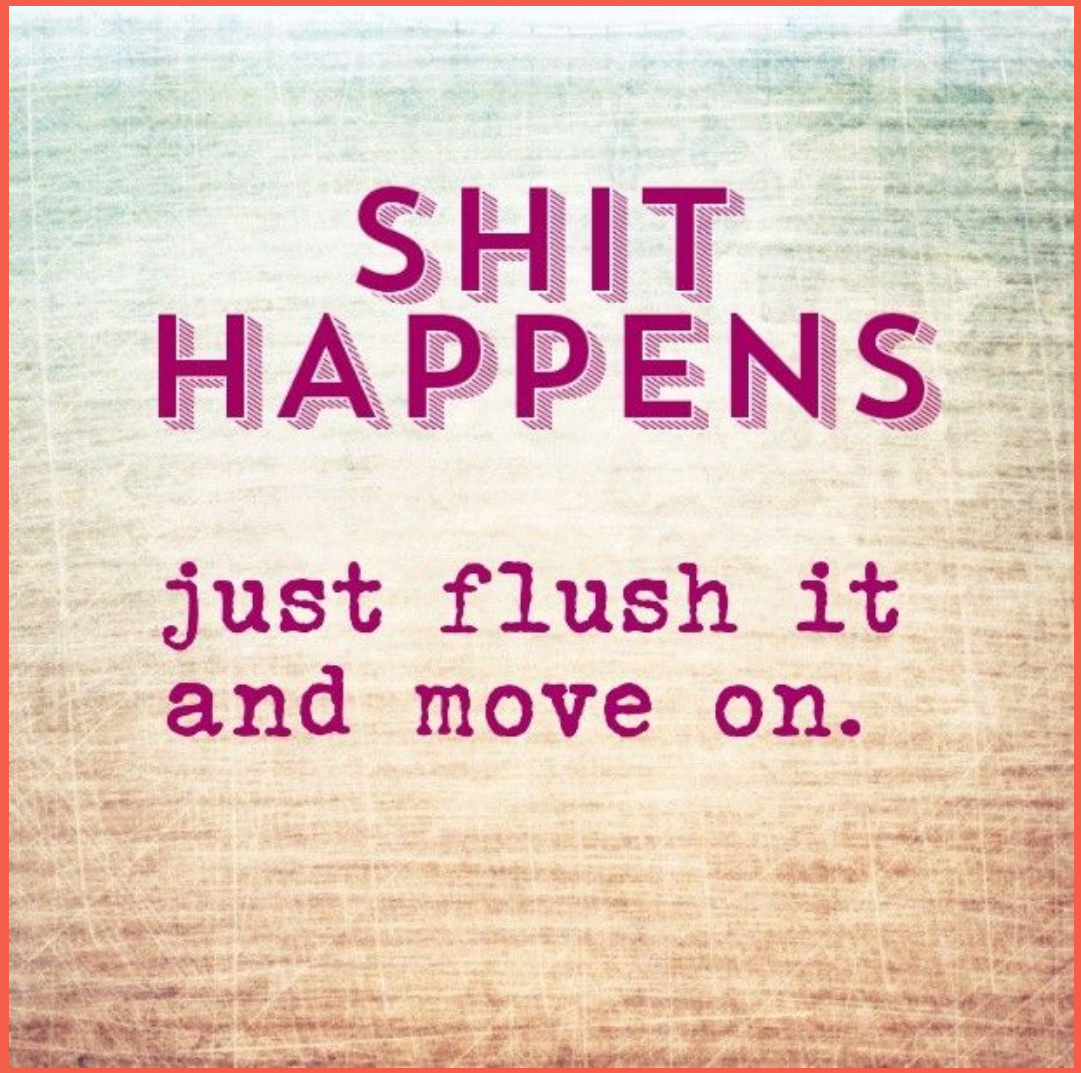
Results

Logistic Regression using TF-IDF Vectorizer was my most accurate model at predicting which subreddit a post came from.

Because both subreddits are giving life advice, it would make since a lot of the words use in both would be the same words.

Since TF-IDF penalizes common words and rare word have more influence it would this would make sense that it would score better.

Further study of unique words and common words to see if model could be improve by weighting or removing words.


SHIT
HAPPENS
just flush it
and move on.