Satisfied or Infuriated?

Can we predict which subreddit a comment was posted in?

Jeffrey Floyd
Data Scientist

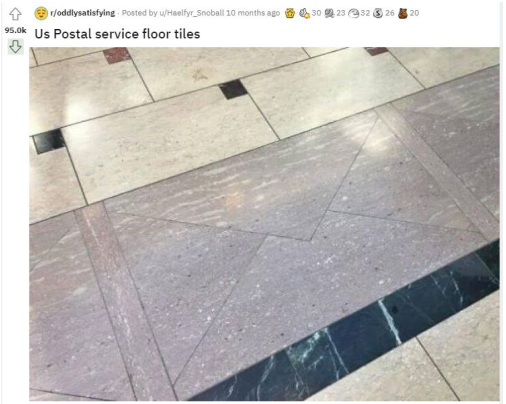
Quick Overview

- Data Collection
  - Subreddits used
  - Pushshift
  - Cleaning
- Data Exploration
  - Key Features
- Modeling
  - Feature Engineering
  - Models Used
- Findings
  - Baseline
  - Model Evaluation
  - Conclusions
- Questions

Subreddits Used

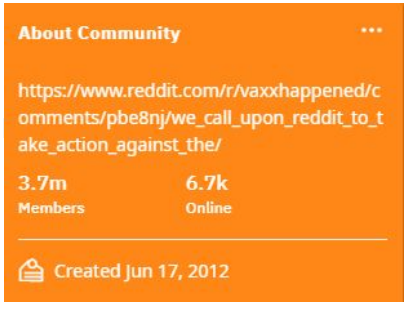
r/oddlysatisfying

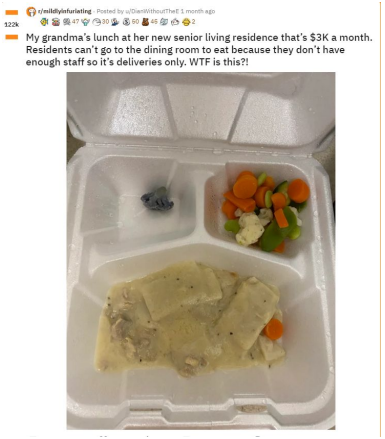
Dedicated to visually or auditorily pleasing images and videos

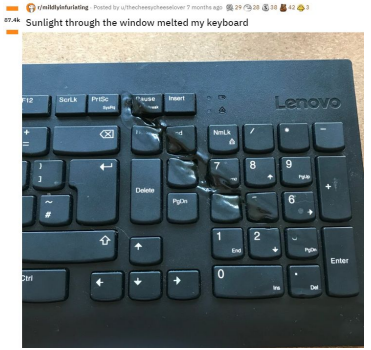


r/mildlyinfuriating

Dedicated to images and videos that trigger your inne OCD or fuel your disdain for people







Data collection and Cleaning

Pushshift.io Reddit API -
The pushshift.io Reddit API was designed and created by the r/datasets mod team to help provide enhanced functionality and search capabilities for searching Reddit comments and submissions.

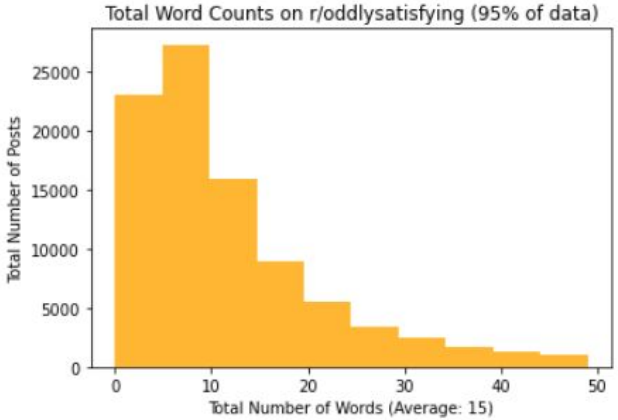
About 100,000 comments were collected from each subreddit
Started with the most recent post as of October 27, 2021 at 23:40:09 UTC
Worked backwards to the next most recent post in batches of 100

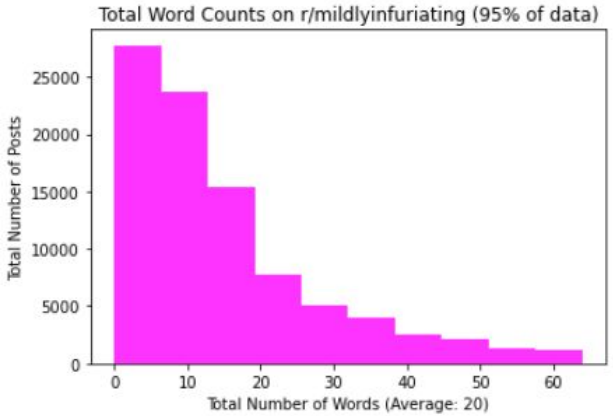
Embedded links had to be cleaned

Removed:
Hyperlinks
[deleted] / [removed] posts

Data Exploration

Average Word Count Per Comment

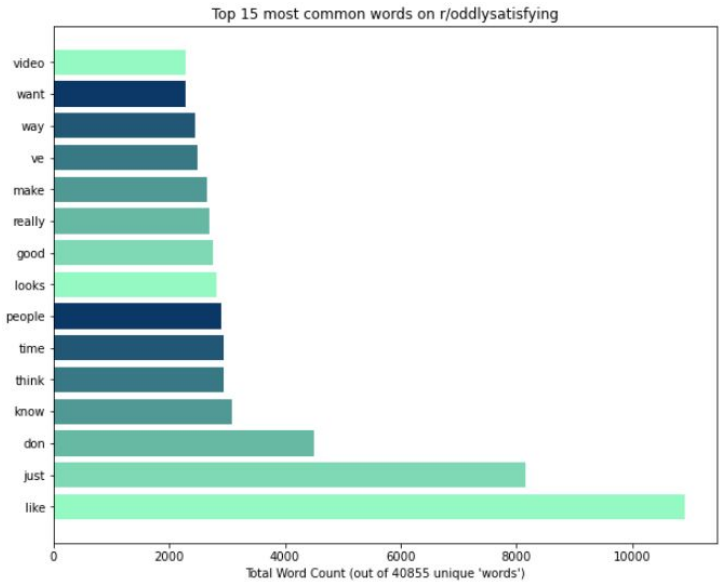


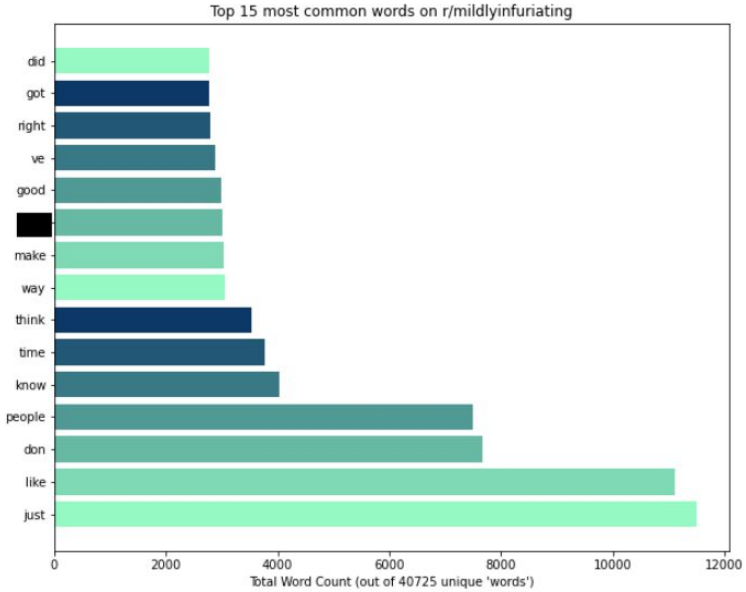


At first glance it looks like there isn't a huge divide in post length but the average is slightly higher on r/mildlyinfuriating

Data Exploration cont.

Most Common Words

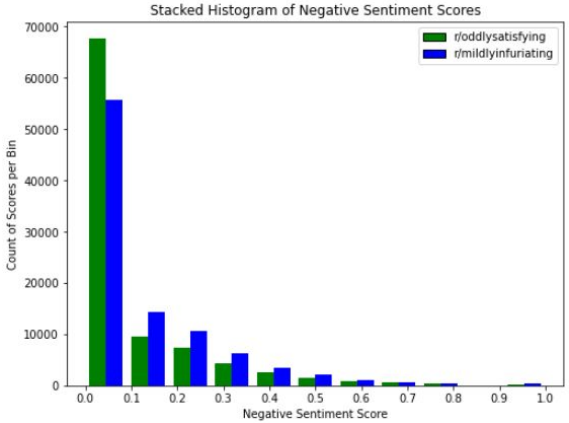


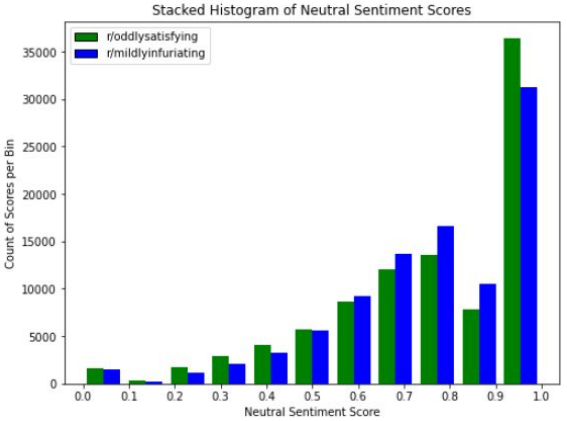


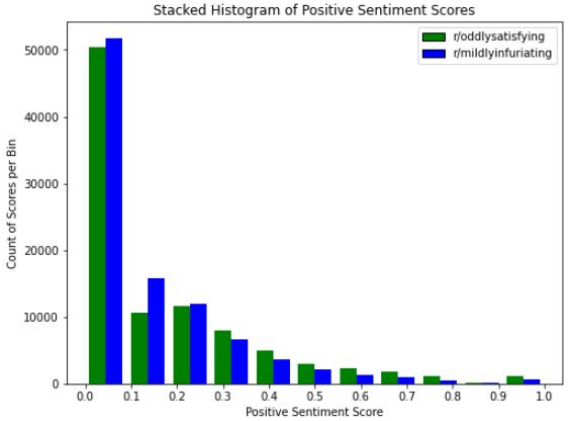
English stop words removed (“the”, “it”, “to”, “and”, etc.)

Data Exploration cont.

Sentiment Analysis







Higher values mean post sentiment more in line with the respective classifier

Modeling

Final comment count at 63,338 (due to computational constraints)

"Word" instances counted with CountVectorizer

00	000	01	02	03	04	05	07	08	10	...	zeros	zikr	zone	zoomed	zucchini	word_count	neg_sent	neu_sent	pos_sent	comp_sent
0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	23	0.000	0.847	0.153	0.5859
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	46	0.000	0.785	0.215	0.8648
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	6	0.000	1.000	0.000	0.0000
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	90	0.074	0.797	0.129	0.7269

Model Types used:

- Logistic Regression
- Random Forest
- XGBoost

Model Performance

Baseline Accuracy: 50.1%

If we guessed just r/oddlysatisfying we would be correct 50.1% of the time

Model Scores:

- ➢ Logistic Regression
  - Training Accuracy: 83.45%
  - Testing Accuracy: 75.81%
  - F1 Score: 76.80%
- ➢ Random Forest
  - Training Accuracy: 82.54%
  - Testing Accuracy: 71.84%
  - F1 Score: 73.26%
- ➢ XGBoost
  - Training Accuracy: 74.59%
  - Testing Accuracy: 72.15%
  - F1 Score: 73.78%

Conclusions

Capped at roughly 76% accuracy for all fitted models

Potential causes:

- Needs more stop word tuning
- Comments on posts that fall outside the subreddit theme
- Users having conversations unrelated to the post through comments
- Single word posts not tracked in our bag of words

Recommendations:

Check out both subreddits!

Questions?

Sources

Reddit logos:
https://logos-world.net/reddit-logo/
https://en.wikipedia.org/wiki/File:DataIsBeautiful.png

Subreddits and Posts:
https://www.reddit.com/r/oddlysatisfying/
https://www.reddit.com/r/mildlyinfuriating/
https://www.reddit.com/r/oddlysatisfying/comments/9rfjyp/my_neighbors_tree_has_the_perfect_fall_gradient/
https://www.reddit.com/r/oddlysatisfying/comments/kl4svw/us_postal_service_floor_tiles/
https://www.reddit.com/r/mildlyinfuriating/comments/ml7643/sunlight_through_the_window_melted_my_keyboard/
https://www.reddit.com/r/mildlyinfuriating/comments/pvz8rm/my_grandmas_lunch_at_her_new_senior_living/

Pushshift: https://github.com/pushshift/api

Two Button Graphic: https://imgflip.com/memegenerator/Two-Buttons