DSI Project 3: Reddit Classification

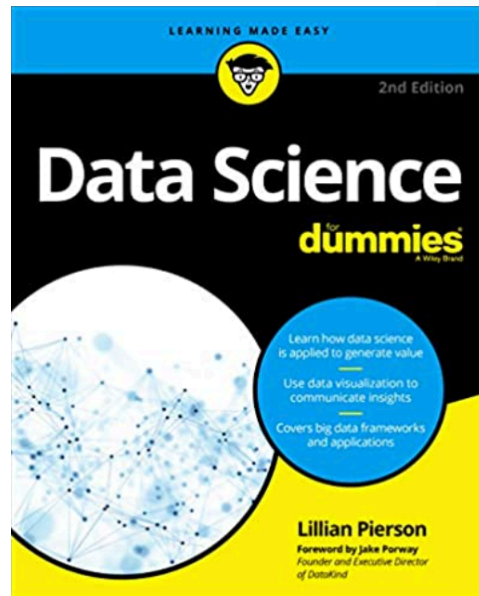**Mok Teck Ching**
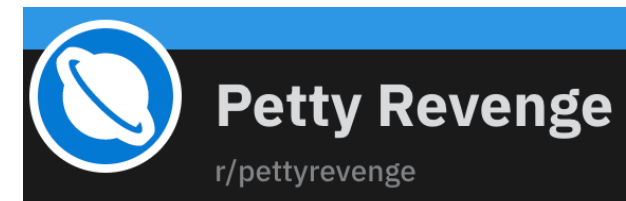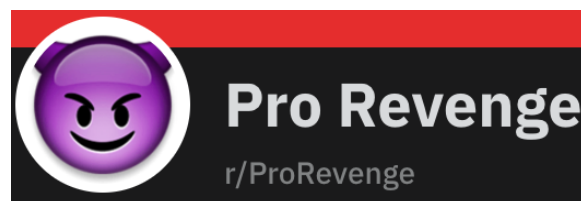
Setting the stage
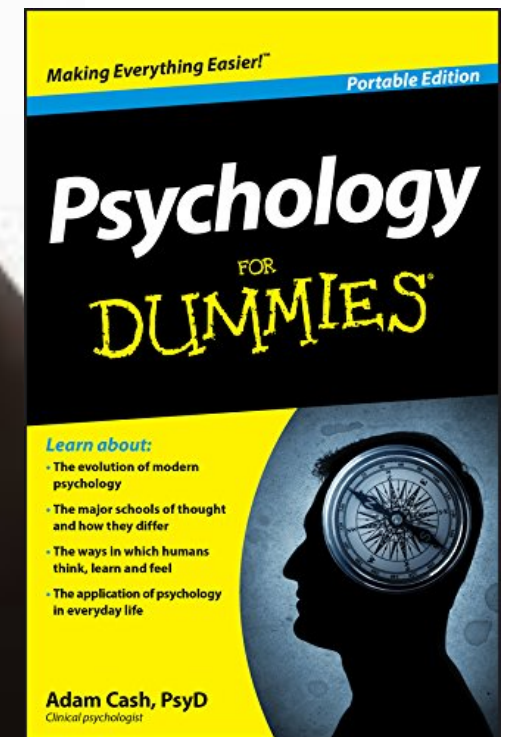
This looks photoshopped

Problem statement

- Gauging maliciousness via language

- Compare 2 subreddits

Problem statement

Use **scraped data** from both these subreddits to build a model that would **determine if a post was less (r/pettyrevenge) or more (r/prorevenge) malicious**

Do a **simple analysis** of any significant words found

Model must have an **accuracy of over 70%**

r/pettyrevenge

Petty Revenge [Join]

r/pettyrevenge

Create Post

🔥 Hot   ✳️ New   📊 Top   This Month ▾   ⋯

↑   21.0k   ↓

Posted by u/Anotherminion1 1 month ago 🏅 🦞3 🎖️2 😎3 & 11 More

Landlord "Always" Keeps Security Deposits? Game On!

Please be gentle, it's my first story.

Many moons ago, after my divorce, I saved up enough money to move from my apartment (a building with 12 units) and buy a house. I was really excited and told my favorite neighbor. He told me not to even bother trying to get my security deposit back because the landlord never returned security deposits.

I turned in my notice. The landlord didn't bother to do a final walk through. After 30 days, I emailed her and asked about my deposit. No response. I filed a small claim court case for the deposit.

This lovely lady's strategy was to hire a lawyer and bump it up to the next higher court. Most people quit after she bumped it up to the next higher court.

💬 588 Comments   🎁 Give Award   ↗ Share   🔖 Save   ⋯

r/prorevenge

Pro Revenge
r/ProRevenge   [Join]

Create Post

🔥 Hot   ✳️ New   📊 Top   All Time ▾   ...

About Community

Have a story of you or someone you know getting back at someone with pro revenge after being wronged? Post it here!
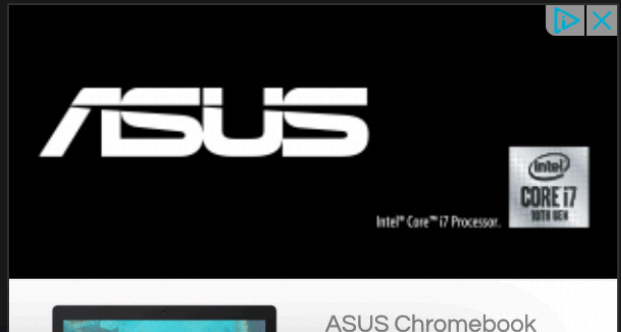
1.1m   **472**
Members   Online

🎂 Created Nov 14, 2012

[Create Post]

Community options   ⌄

⬆️ **69.9k** ⬇️

Posted by u/[deleted] 1 year ago 🗃️ 🌐 3 🏆 8 🦅 Ⓢ 14 & 2 More

Coworker tried to get me fired over breast implants, so I pulled a reverse uno card.

TLDR: Coworker harassed me over having implants (which I only got because I had cancerous cells), I called her out, she reported me to HR for calling her out, I got a bunch of people to write statements of all the times she harassed me and she got fired instead. Also I've been told that isn't what a reverse uno is but I can't change it.

(EDIT: Just want to add that regardless of my reasons for getting implants, I still should not have been harassed over them. It just happened to be relevant here because it made Jill look like a real dick which is probably why she reported me.)

---

r/ProRevenge Rules

1. Posts must be stories of revenge   ⌄

2. Posts must be feasibly possible   ⌄

Preprocessing

[removed] posts were filtered out during the scrape

```python
for submission in data:
    try:
        if submission["selftext"] != "[removed]":
            subCount+=1
            aggregate.append(submission)
        else:
            pass
    except:
        pass
```

'Selftext' Rows with null values were dropped

```
1  # checking for null values
2
3  sorted_df.isnull().sum()

title        0
selftext     8
subreddit    0
dtype: int64
```

	title	selftext	subreddit
408	Does anyone know how to make an account on bad...	NaN	pettyrevenge
872	my boyfriend (gay) dumped me, so I catfished h...	NaN	ProRevenge
1020	Client stiffed me on payment after I SEO optim...	NaN	ProRevenge
1068	K	NaN	ProRevenge
1082	Identity thieves phone Njm	NaN	ProRevenge
1158	How I went "Heathers"	NaN	ProRevenge
1447	Pp touch	NaN	ProRevenge
1580	Oof, I'm pretty sure my text text text text te...	NaN	ProRevenge

posts were further processed

step 1. Remove links

step 2. Remove punctuation and other weird characters

step 3. Remove stop words

step 4. Remove words with less than 3 letters

step 5. Lemmatise tokens

Word Count < 100 dropped

	title	selftext	subreddit	text	wordcount
7	e	e	pettyrevenge		0
10	Vui thôi đừng vui quá :))	Funny dog haha	pettyrevenge	funny haha	10
303	Would pay good money for this	Just saw this. There's only the one ex I have ...	pettyrevenge	would good money mind	21
341	Fart story	My brother blamed me for farting at school, so...	pettyrevenge	fart story brother blamed farting school night...	64
522	(Request) I have a person phone number. I real...	Title	pettyrevenge	request person phone number really like anythi...	59
528	CALL THIS GUY AND DONT STOP	Hey guys so this guy was being mean to my favo...	pettyrevenge	call dont stop guy mean favorite streamer call...	98
582	Please sign	https://www.change.org/p/reddit-have-elon-mus...	pettyrevenge	please sign	11
717	Shadow	I think yes!!	pettyrevenge	shadow think	12

Count Vectoriser

TF-IDF Vectoriser
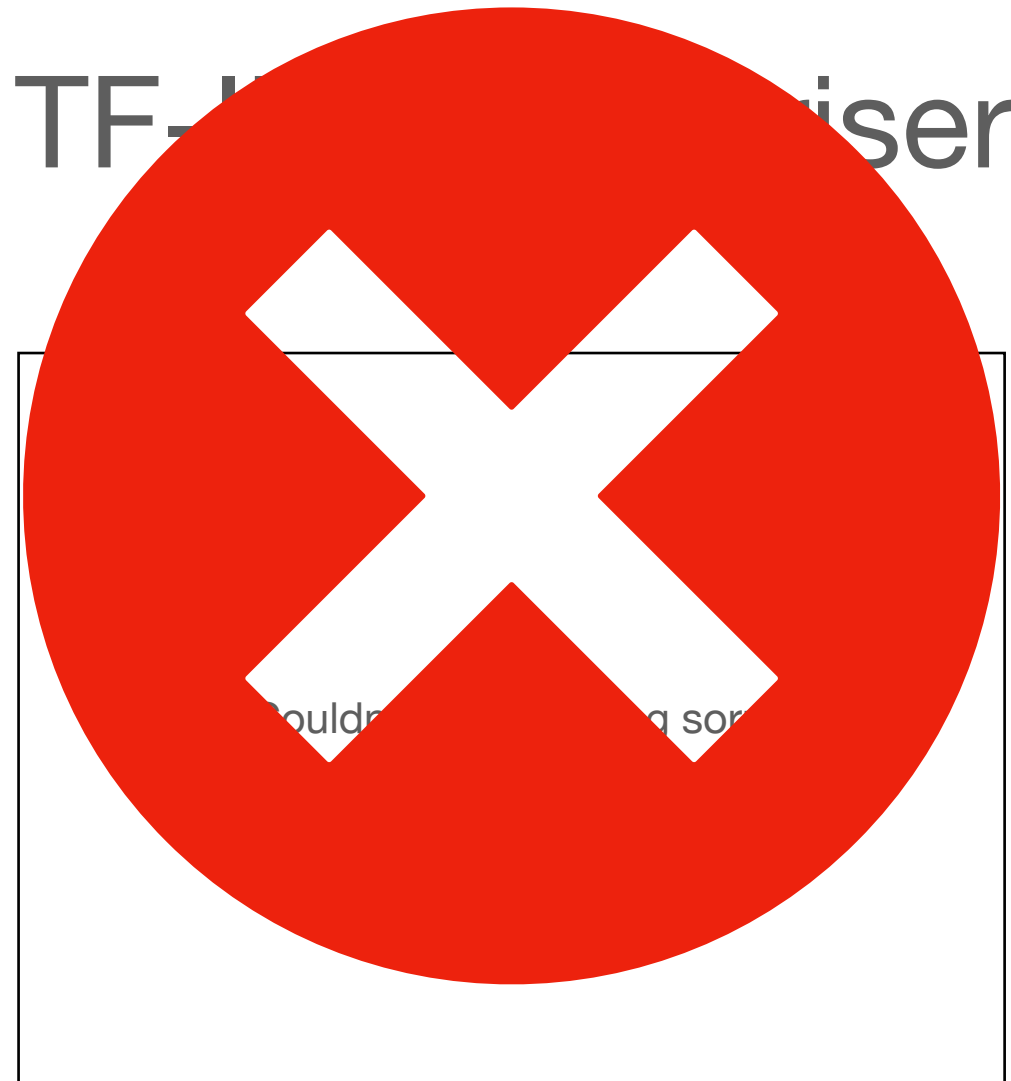
EDA

r/pettyrevenge   r/ProRevenge
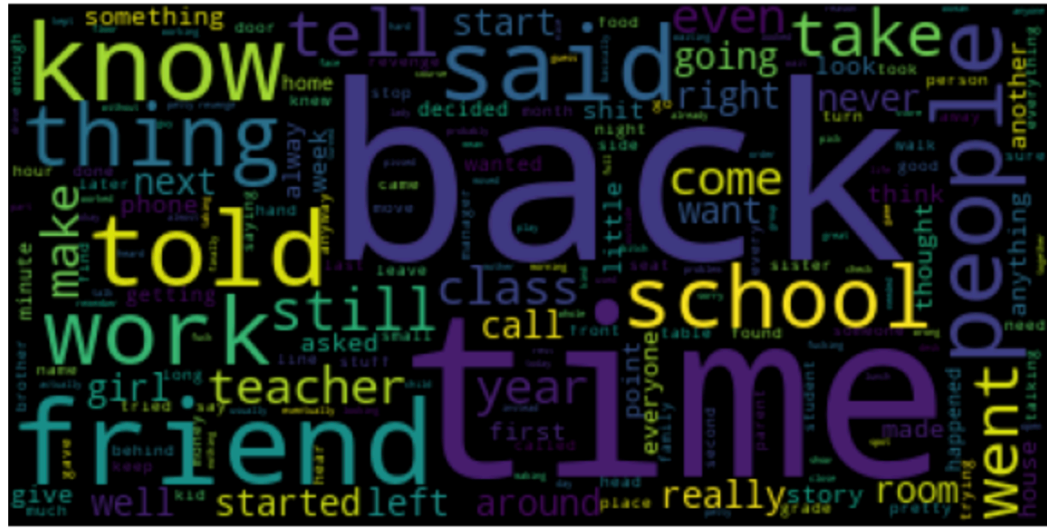
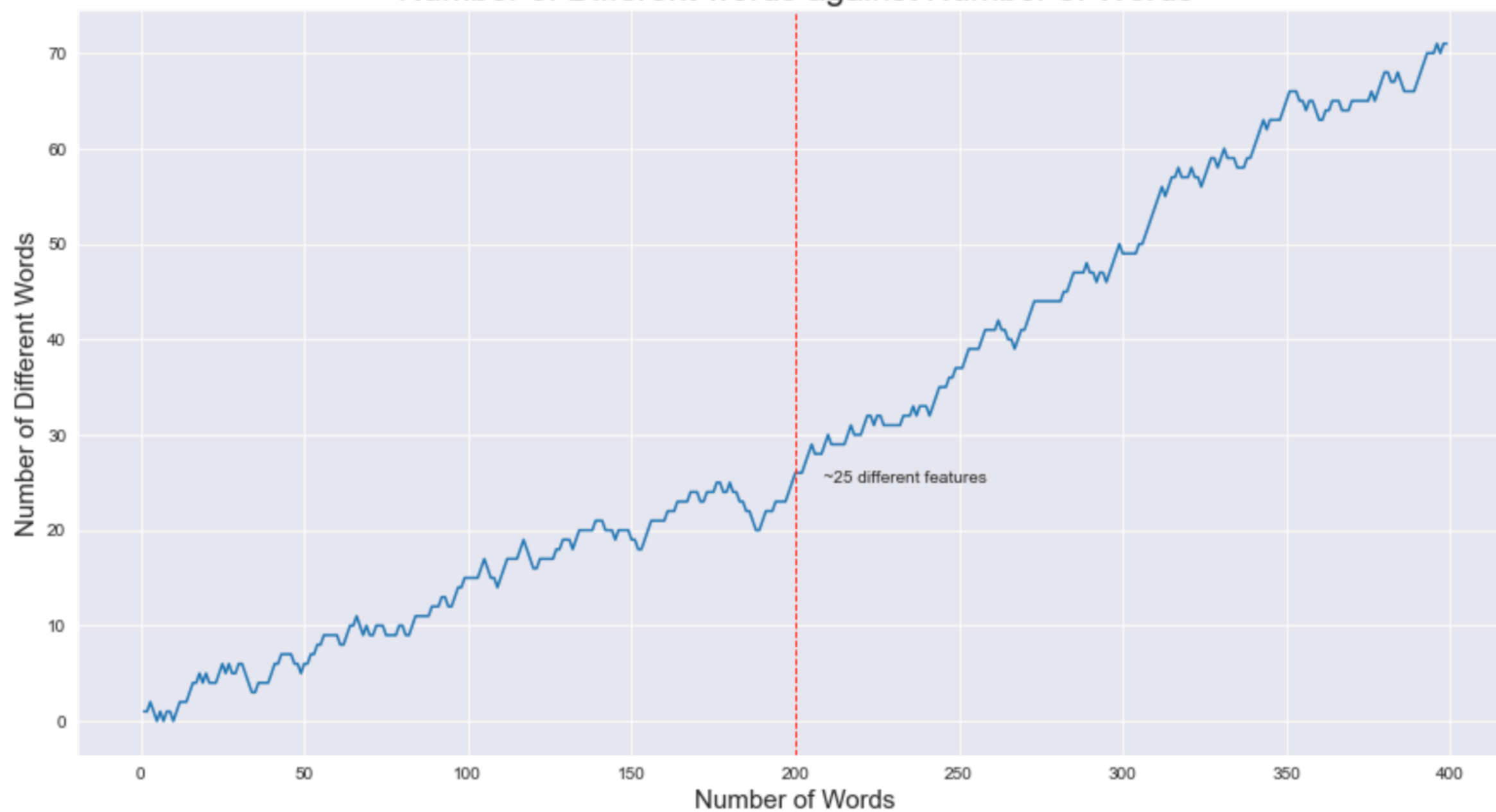# Data Cleaning

Number of Different words against Number of Words

~25 different features

"petty" and "pro" dropped

```python
1  # removing "petty" and "pro" from the dataset.
2  for i in ["petty", "pro"]:
3      try:
4          final_df.drop(columns = [i], inplace = True)
5      except:
6          pass
```

CountVectorized Scores of top 20 unique words from both subreddits
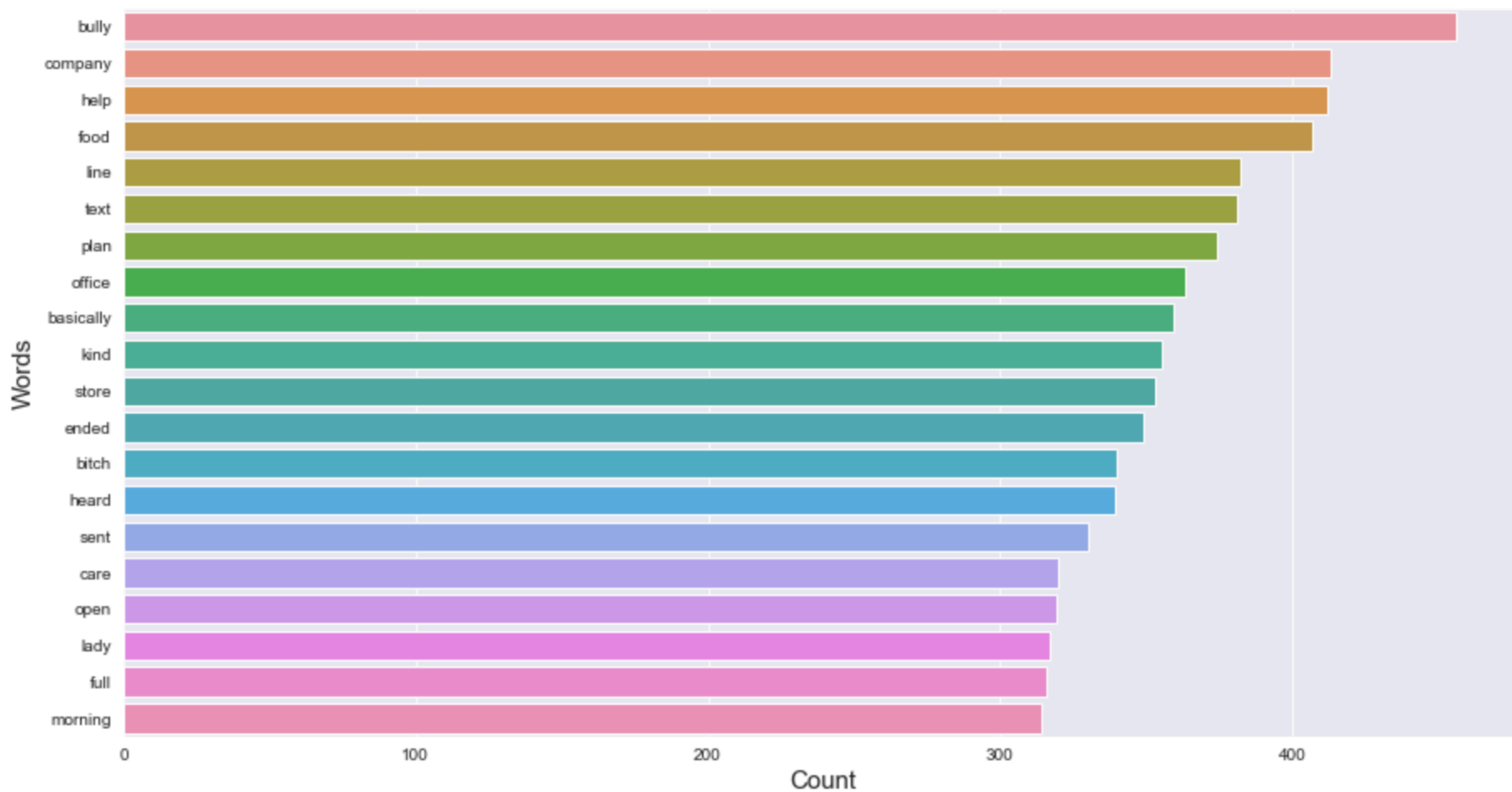
Model Building

Multinomial Naive Bayes	K Nearest Neighbors	Random Forest
MultinomialNB( )	KNearestNeighborsClassifier ( )	RandomForestClassifier ( )
train score: 0.76 test score: 0.66 Con matrix: True Negatives: 147 False Positives: 63 False Negatives: 75 True Positives: 126	train score: 0.70 test score: 0.52 Con matrix: True Negatives: 181 False Positives: 29 False Negatives: 170 True Positives: 31	train score: 0.77 test score: 0.71 Con matrix: True Negatives: 148 False Positives: 62 False Negatives: 59 True Positives: 142

Multinomial Naive Bayes	K Nearest Neighbors	Random Forest
MultinomialNB( )	KNearestNeighborsClassifier ( )	RandomForestClassifier ( )
train score: 0.76 test score: 0.6? Con ... True Negatives: 147 False Positives: 63 False Negatives: 75 True Positives: 126	train score: 0.70 test score: 0.5? Con ... True Negatives: 181 False Positives: 29 False Negatives: 170 True Positives: 31	train score: 0.77 test score: 0.71 Con matrix: True Negatives: 148 False Positives: 62 False Negatives: 59 True Positives: 142

Model Optimization

Random Forest

RandomForestClassifier ( )

train score: 0.7457351746547523
test score: 0.7274939172749392

Con matrix:

True Negatives: 154
False Positives: 56
False Negatives: 56
True Positives: 145

Best parameters set: {'max_depth': 2, 'n_estimators': 166, 'random_state': 42}

Discussion

- The words "Bully" and it's deriatives, as well as the word "abuse" hint that presence of hostile intent from the offending party could cause further escalation.

- Place of work vs school was also another significant theme, judging by the words "company" and "office". Finally, there are some words that allude to forward planning in carrying out revenge, e.g. "plan" and "idea".

- From this, it can be theorized that an increase in maliciousness can be characterized by 1) the presence of a hostile party, 2) issues in a workplace/school or 3)forward planning.

- One limitation of this model is the fact that maliciousness is not a binary variable. While the task was to compare only 2 subreddits, it could be argued that creating a "scale" of malicousness by using several other "revenge" subreddits, such as r/nuclearrevenge or r/maliciouscompliance, would serve to better highlight the nuances in the text.