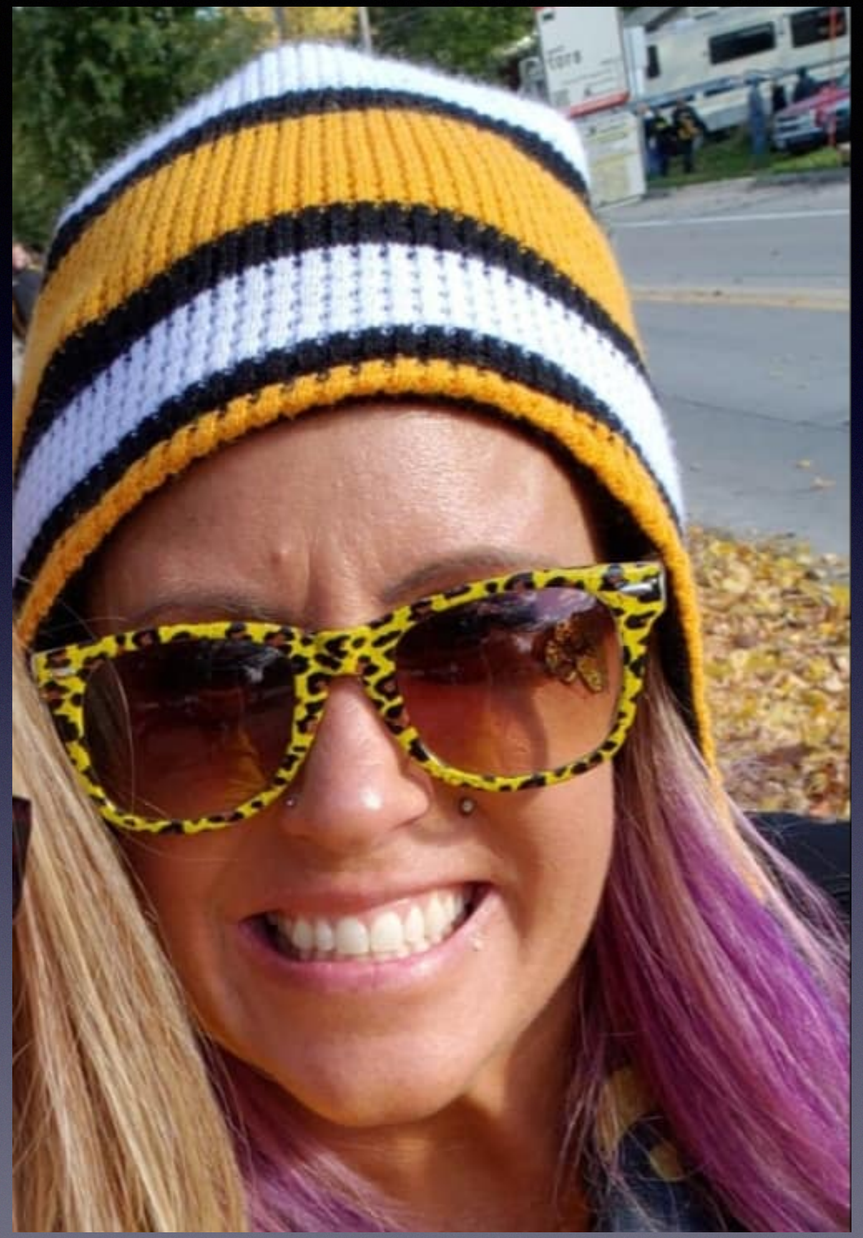
Oh the Horror!

NLP Analysis on Reddit Comments

This is my friend Kari and she loves a good horror movie.

She would like to find a good place to find some new horror movies. But, she is tired of people ragging on movies and directors. She wants to find a place to get ideas for movies to watch, not to hear about movies not to watch.



A tale of two subreddits…

- I found two horror movie genre reddits with a decent following:

DREADIT
THE HORROR REDDIT

- r/horror is All Things Horror
- r/HorrorMoviesONLY is Horror Movies ONLY!

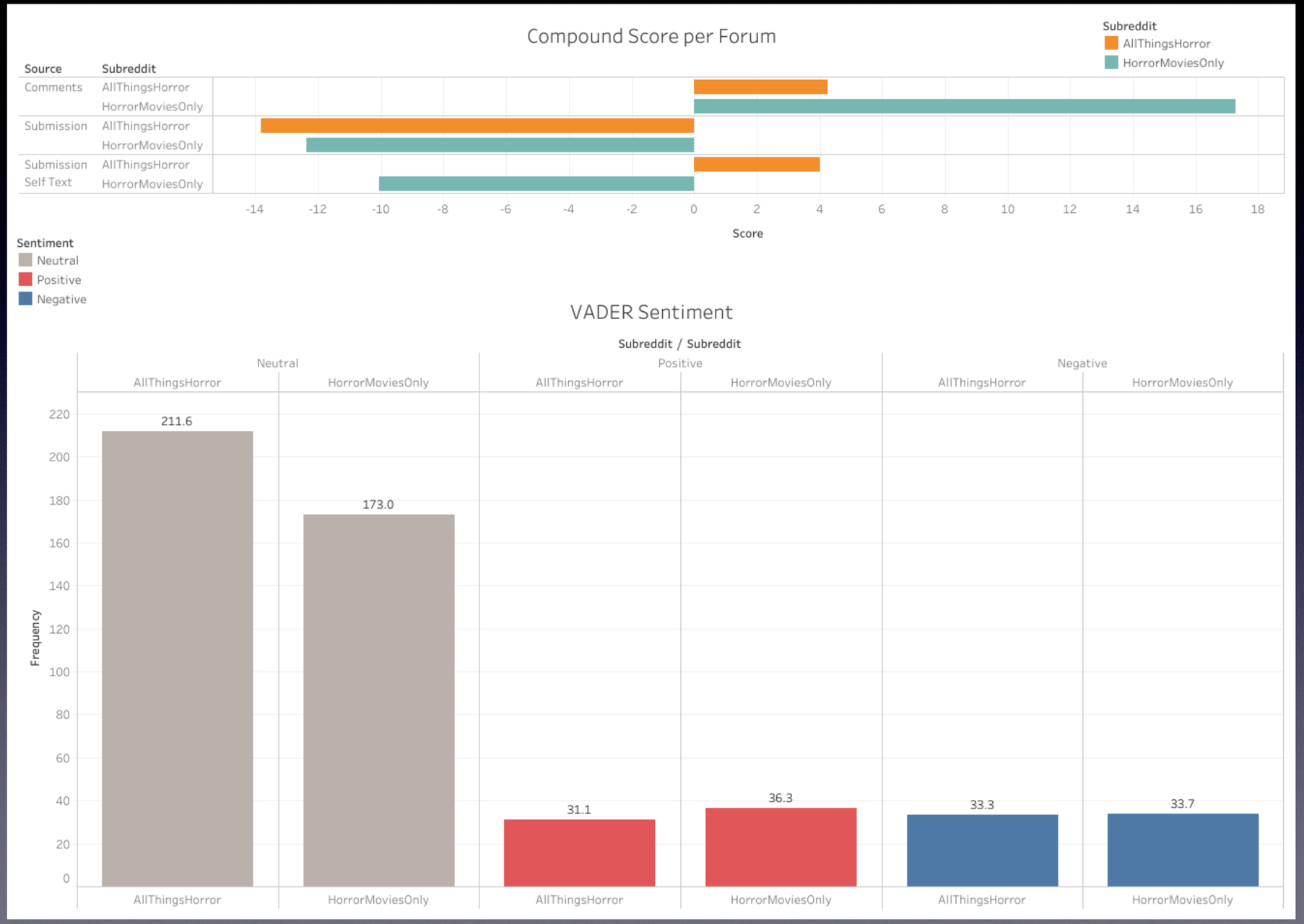
(All Things Horror does include other mediums besides movies but I wanted to compare a larger forum with a smaller participation rate.
And, All Things Horror mainly posts about movies.)

How to give Kari a good review site?

I chose two main indicators for analysis…

How much gore do people like?

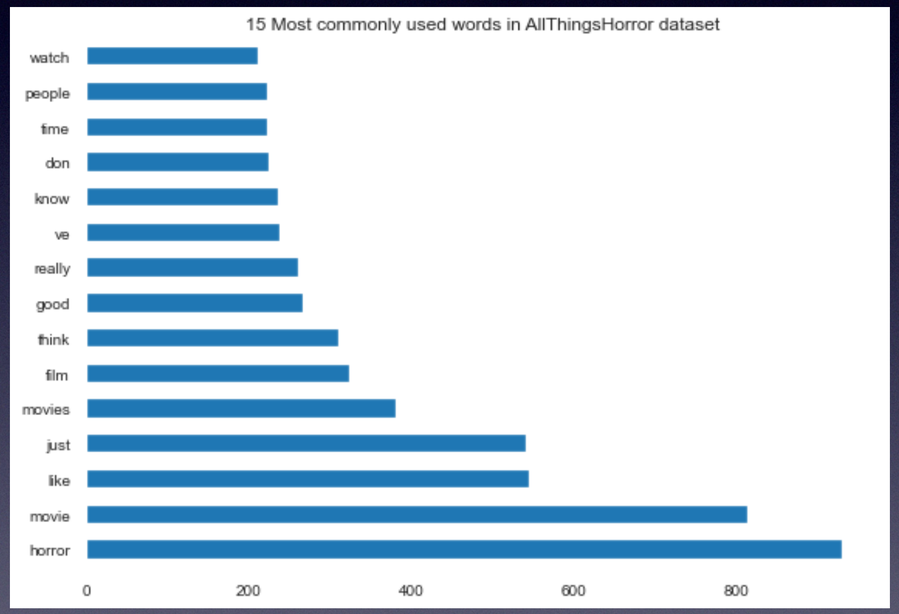
- The first one is a sentiment analysis.
- I combined all of my values into one large database, which included individual rows for the “subtext” field.
- I used VADER to assign a sentiment to each statement.

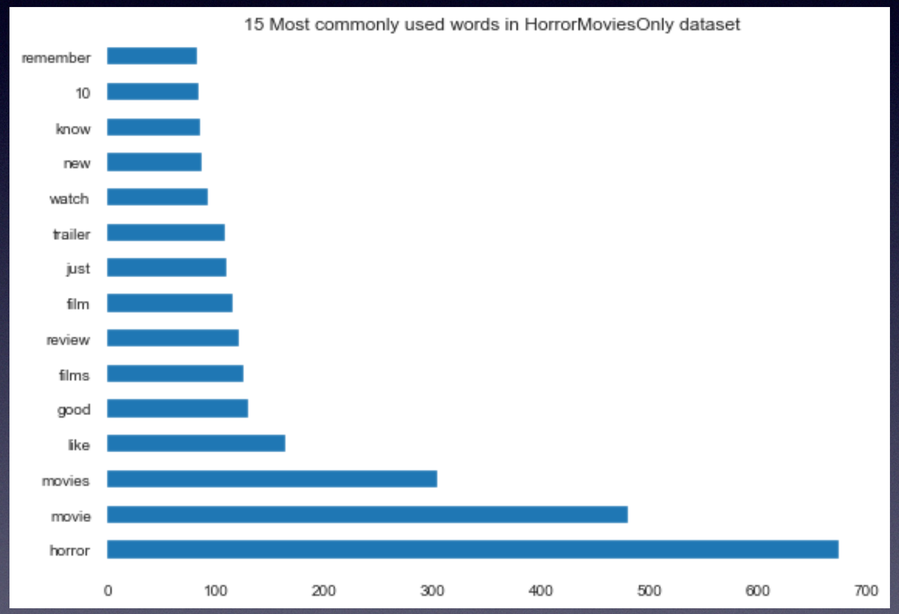


- VADER found r/HorrorMoviesOnly to have a lot of positive and a lot of negative scores.
- r/AllThingsHorror had a higher “neutral” score.
- What’s wrong with this analysis?

Sentiment for Horror Genre is tricky…

- Here are the top 15 words for each site:





- Since horror was the top word used by both sites, and VADER sets horror with a very negative sentiment score, I needed to remove it from the lexicon.

Removed the following instances of “horror” and “scare’” along with their scores/weights:

```
horrific  -3.4  0.91652
horrifically -2.9  0.7
horrified -2.5  0.92195
horrifies -2.9  1.22066
horrify -2.5  0.67082
horrifying  -2.7  0.9
horrifyingly -3.3  0.9
horror  -2.7  1.1
horrors -2.7  0.64031

scare -2.2  0.87178
scarecrow -0.8  0.9798
scarecrows  -0.7  1.1
scared  -1.9  0.7
scaremonger -2.1  0.53852
scaremongers  -2.0  1.0
scarer  -1.7  0.78102
scarers -1.3  0.9
scares  -1.4  0.4899
scarey  -1.7  0.64031
scaring -1.9  1.22066
scary -2.2  0.87178
```

Whose Gore is it?

- The second analysis I did was on a model to predict which site a post came from.

BEST Forum Model:

Naive Bayes with TfidVectorizer.

Score: 81%

Best parameters:

Setting terms to ignore words that appear in less than 2 posts and words that occur in more than 80% of the posts.

It seems there was a sweet spot of commonly used words because there were unique words specifically for each forum.

There was an advantage to viewing words in pairs of 2.

English stop words helped.

BEST Sentiment Model:

Vote Ensemble with DecisionTree, Boosters and Count Vectorizer.

Score: 84%

Best parameters:

Setting terms to ignore words that appear in less than 2 posts and words that occur in more than 90% of the posts.

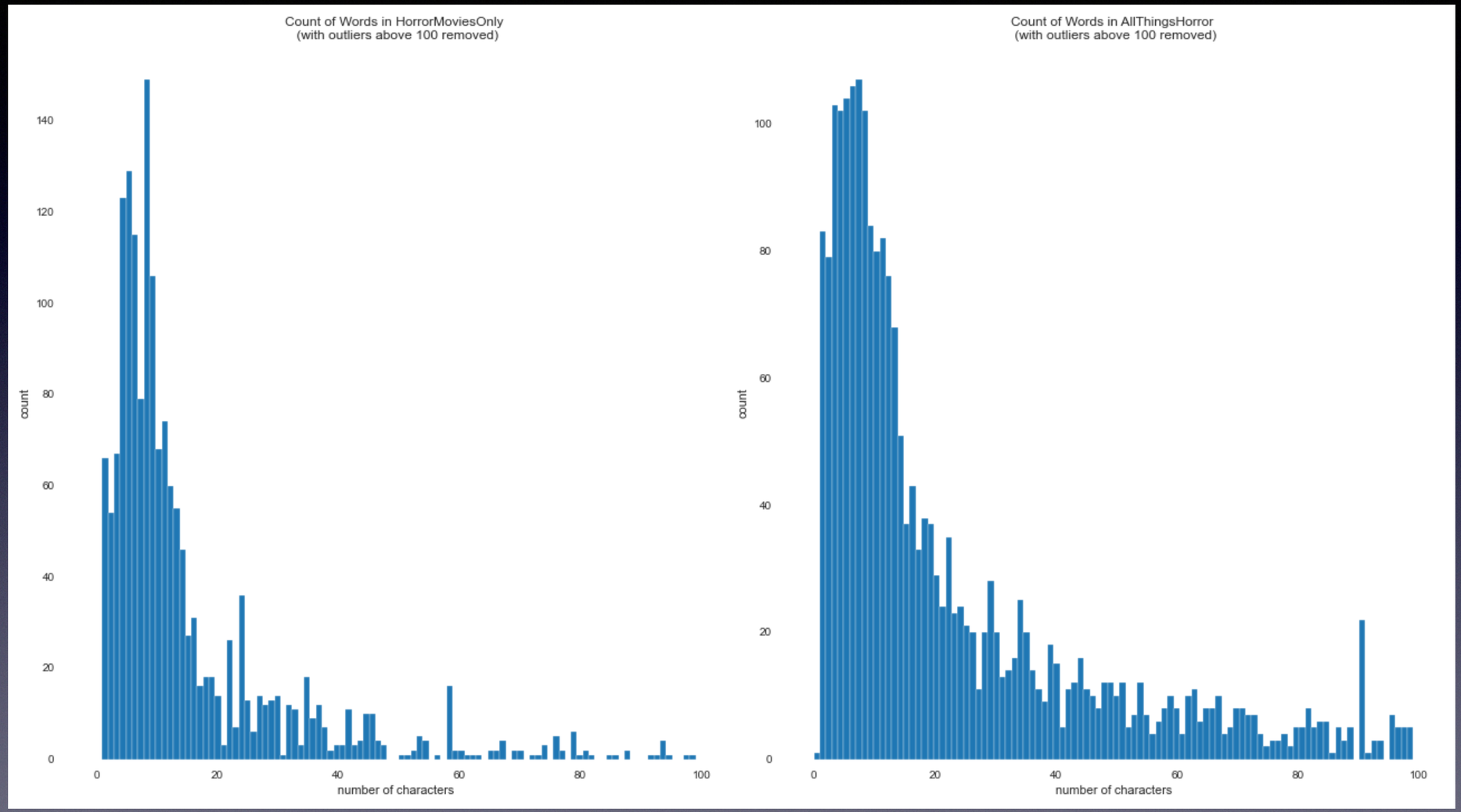
It seems there was a wide variety of words across all posts that was harder to pin to one sentiment vs. another (i.e. Lexicon)

There was an advantage to viewing words in pairs of 2.

Not using stop words in this model helped.

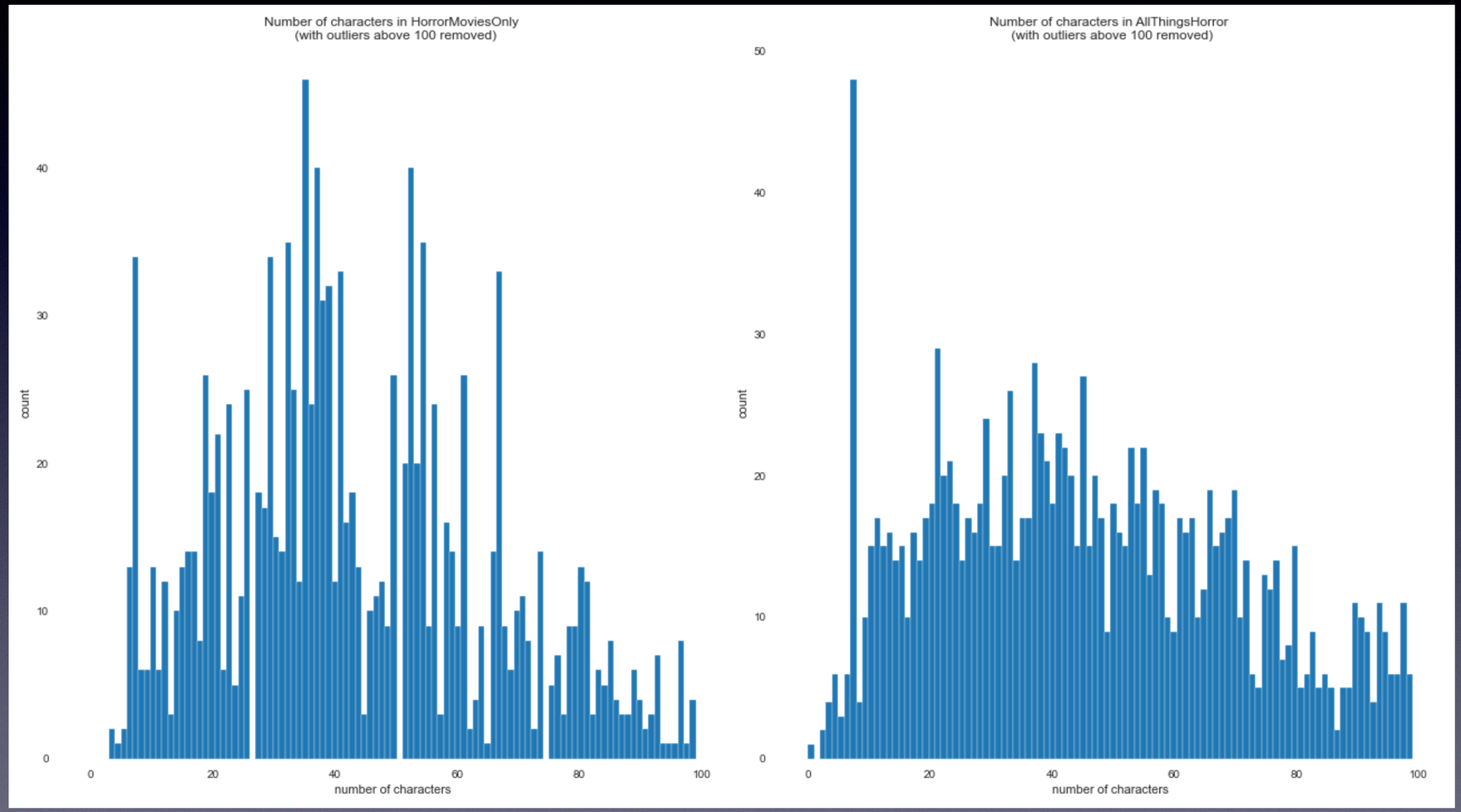
Which site should Kari go to?

Recommendation: All Things Horror



- Looking at the count of words in both forums, r/AllThingsHorror has more data. But, spikes from 0-10 words per post. While r/HorrorMoviesOnly spikes from 10-20 words per post.

- I believe r/HorrorMoviesONLY is more subjective as a small, subject focused site and prone to over exaggerated feelings. This is again demonstrated in the number of characters. I believe the large variability in r/HorrorMoviesOnly is the result of wordy, highly subjective posts.



- r/AllThingsHorror is normally distributed around 38 characters BUT r/HorrorMoviesOnly has spikes at 10, 35, 50 and 70.

How to improve the chances of finding Kari a movie she wants to see?

Improve Sentiment

- Garbage in, garbage out. There is no lexicon for the horror genre.
- There is a need for this though. Examples for recommender systems are:
    - “A new detective show?“
    - “Books similar to Steven Kings’ ?“
    - “I want to listen to a good, dark ambient playlist…“

- There is an extended model that I could build. There are several discussions of this on the internet.

1. I would scrape genres from IMDB.

2. Match those genres to the movies in the reddit comments.

3. Use NLP to decide which forum matches more closely with the genre Kari is looking for.

This would direct Kari to specific forums that talk positively about her favorite genre: slasher films.

References:

- Github: Pandas Profiling: (https://github.com/pandas-profiling/pandas-profiling)
- Predicting Movie Genres using NLP - An Awesome Introduction to Multi-Label Classification: (https://www.analyticsvidhya.com/blog/2019/04/predicting-movie-genres-nlp-multi-label-classification/)
- Images: IndieWire, Smashing Magazine, and ClickWallpapers.