Reddit Post Classification

Adrian Chapman

Classifying Posts to Improve Communities on Reddit

A Star Wars Subreddit for Everyone!

Saltier Than Crait is a community for those who are critical of the recent new Star Wars revival from Disney and wish to have intelligent, respectful discourse about it.
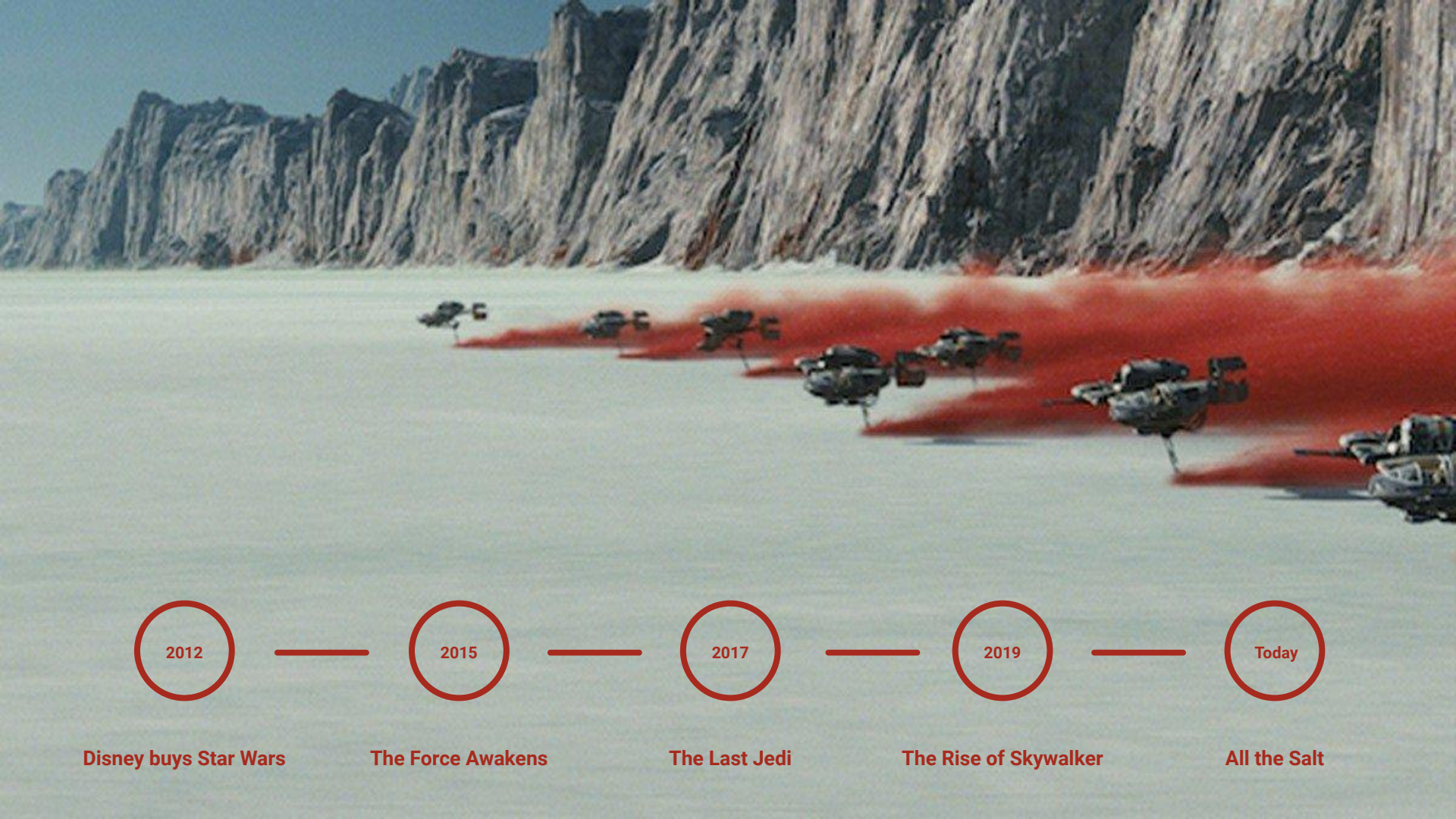
69.1k Salt Miners

Created Dec 16, 2017

Star Wars Cantina: A place to discuss Star Wars with a huge emphasis on positivity. Criticism of any content is allowed, but do so respectfully towards the creators as well as the fans. The Cantina is here for us to celebrate what we love about Star Wars.

58.4k Cantina Regulars

Created Mar 2, 2018

2012 — 2015 — 2017 — 2019 — Today

Disney buys Star Wars   The Force Awakens   The Last Jedi   The Rise of Skywalker   All the Salt

It's just nice to see that there are still people who enjoy the franchise without bashing on each other

Me finally finding a wholesome Star Wars community after all the fandom drama in recent years

Chewie, we're home

Problem Statement

- Reddit is a community driven message board with many subreddits that users can subscribe to. Each subreddit is highly moderated and has its own rules, community, and feel.
- If a post does not meet the expectations of the subreddit community, moderators may suggest posting in a different subreddit, may remove the post, or may even ban the user.
- Can we use Saltier Than Crait historical content to train a classification model to help Star Wars Cantina moderators discover posts too negative for the Cantina community

# Exploratory Data Analysis

Data Cleaning

- Used PRAW to pull data
- ~ 1,000 posts per subreddits
- Removed hyperlinks and punctuation
- Removed stopwords
- No lemetizing or stemming
- Kept short posts (1-3 words)

Word Count

After combining the title and body of each post into a single feature, CountVectorizer was used to explore the most common words by subreddit.

Additional features were created for total word count and character length for each post.

The large difference between average word count between subreddits indicates that word count may be a useful feature to include in the modeling process.

Most Common Words: Star Wars Cantina

Most Common Words: Saltier Than Crait
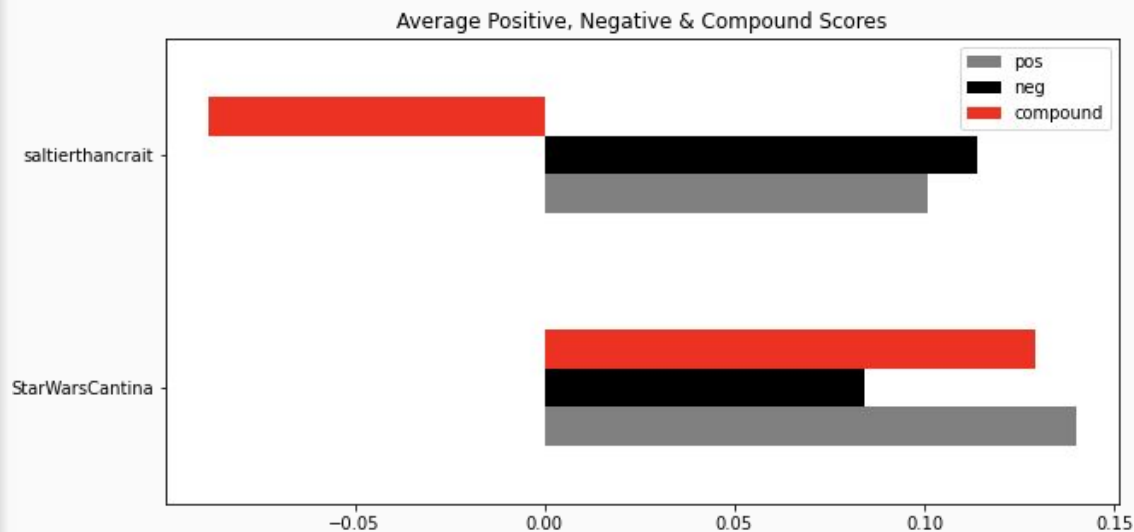
Average Word Count by Subreddit
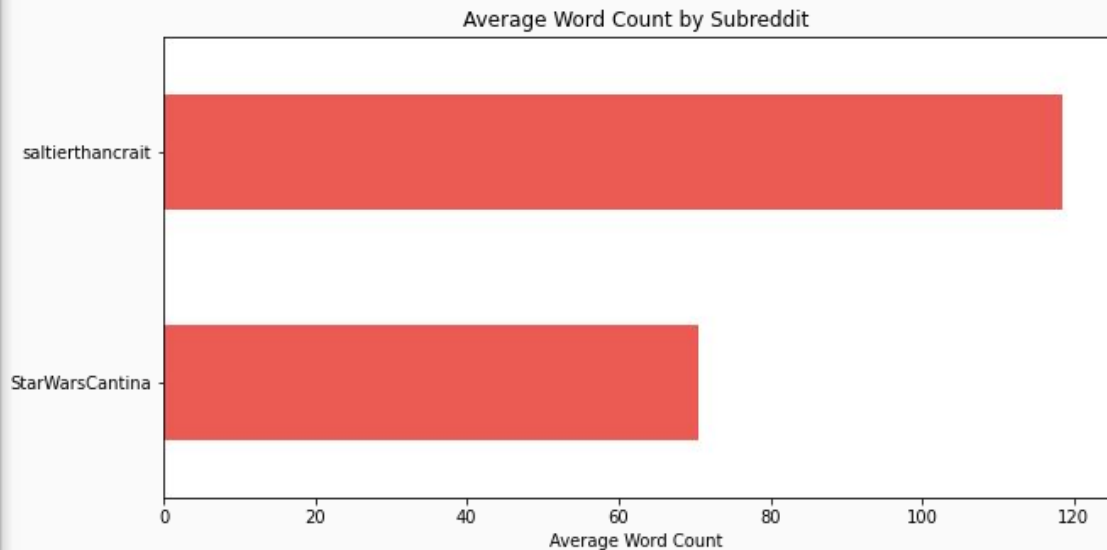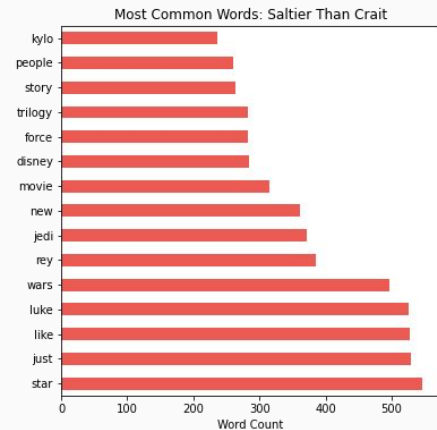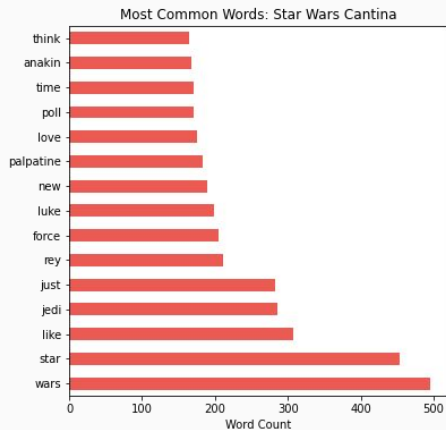
Sentiment

The VADER sentiment analysis tool was used to determine if there was differences in sentiment between the two subreddits.

Saltier Than Crait indeed is more negative on average than Star Wars Cantina.

The VADER scores were also kept for modeling purposes.

Average Positive, Negative & Compound Scores

neg	neu	pos	compound	features	target
1.000	0.000	0.0	-0.3612	Ridiculous.	1
0.722	0.278	0.0	-0.6369	It *was* a trap :(	1
0.674	0.326	0.0	-0.7351	RIP this poor poor cast	1
0.669	0.331	0.0	-0.8442	The People Who Killed Star Wars	1
0.629	0.371	0.0	-0.5255	I had a goddamn PLAN!	1

Model Evaluation

Model Evaluation

While working with Naive Bayes, Count Vectorizer marginally outperformed TFIDF, but for the tree based models TFIDF was superior.

After optimizing for TFIDF hyperparameters on some test models, a vecotrized text dataframe was created and variables for sentiment (from VADER) and word count were added before train/test split.

All models used gridsearch to optimize hyperparameter performance.

	Train Accuracy	Test Accuracy	Specificity	Recall
Naive Bayes with Count Vectorizer	83.8%	75.6%	72.3%	78.8%
Naive Bayes with TFIDF	85.9%	73.1%	64.1%	82.3%
Decision Tree with Gridsearch	90.9%	69.1%	76.3%	62.0%
Random Forest	99.7%	73.3%	71.8%	74.8%
Extended Trees	99.7%	75.2%	69.9%	80.4%
Bagging with Decision Trees	97.8%	70.3%	75.1%	65.6%
Ada Boost	85.4%	75.3%	80.7%	70.0%
Gradient Boost	91.7%	78.3%	81.1%	75.6%
VOTE: Extended, Ada, Gradient	93.2%	78.0%	80.7%	75.2%

Confusion Matrix

The best performing model was a voting classifier combining Extended Trees, Ada Boost, and Gradient Boosted Trees.

- Accuracy: 77.9%
- Specificity: 80.7%
- Recall: 75.2%
- Precision: 79.6%

This was selected as the production model as it greatly outperformed the baseline model accuracy of 50.1%, and a high Specificity.

Confusion Matrix for Voting Classifier
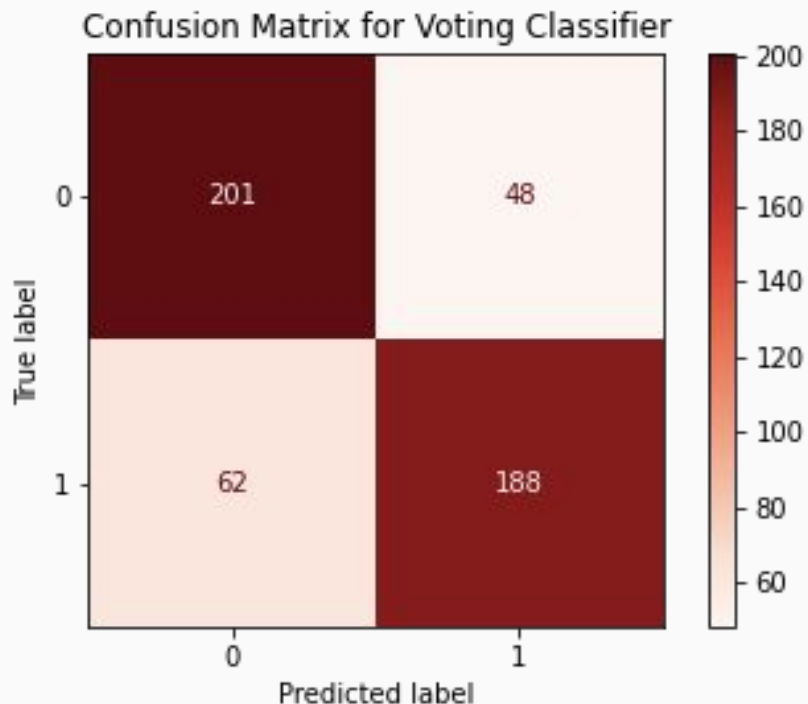
Important Features

Examining the Importance Scores from the Ada Boost model gives insights into the features most useful for modeling.

The "compound" score from the VADER model was the single most useful feature.

Words with Highest Importance Score from Ada Boost
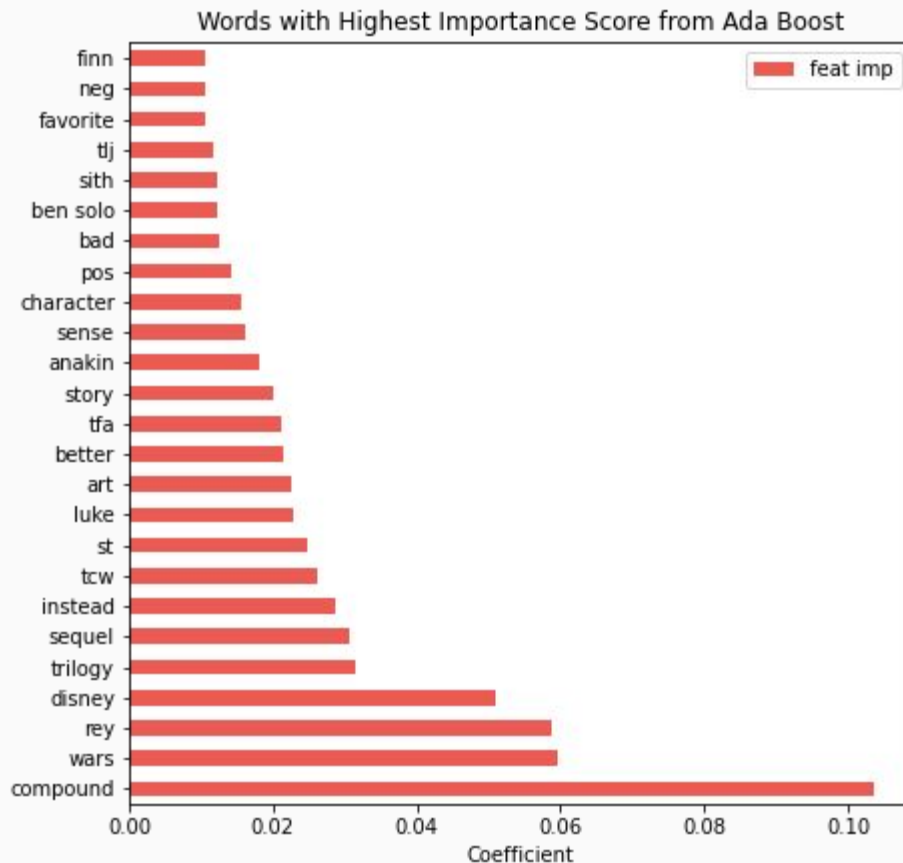
Future Research & Conclusion

Misclassified Posts

Exploring misclassified posts can help aid with model improvement, especially for false positives.

Many false positives are either too short to be sufficiently classified.

Another subset are being flagged as having negative language. These post are probably best reviewed by a human moderator.

False Positives

	features	feature_length	feature_word_count	compound	neg	neu	pos	target	predictions
24	don't know if this is allowed here so please t…	6477	1082	-0.9873	0.089	0.849	0.062	0	1
31	friendly reminder	18	2	0.4939	0.000	0.238	0.762	0	1
32	an autumn baby hbd mark hamill	35	6	0.0000	0.000	1.000	0.000	0	1
39	interesting that the uproar over whether or no…	237	40	0.2023	0.086	0.788	0.126	0	1
42	i dont want to see a star wars movie i want to…	1033	187	-0.8409	0.146	0.768	0.086	0	1
53	live fight another day boys live to fight anot…	57	10	-0.6369	0.394	0.606	0.000	0	1
68	at st goals	12	3	0.0000	0.000	1.000	0.000	0	1
84	about the sequel trilogy	26	4	0.0000	0.000	1.000	0.000	0	1
89	dude	5	1	0.0000	0.000	1.000	0.000	0	1
96	cantina patrons i humbly present to you the st…	10078	1841	-0.7289	0.070	0.855	0.075	0	1

False Negatives

	features	feature_length	feature_word_count	compound	neg	neu	pos	target	predictions
0	regardless of whether it was necessary what gr…	121	14	0.0000	0.000	1.000	0.000	1	0
2	perhaps we treated vector prime too harshly	44	7	0.0000	0.000	1.000	0.000	1	0
13	ahmed best sharing some black jedi love today …	76	13	0.9538	0.000	0.316	0.684	1	0
23	somebody please explain why lasers in space no…	78	12	0.3182	0.000	0.827	0.173	1	0
50	do you believe that george lucas himself holds…	154	20	-0.8074	0.338	0.595	0.067	1	0
55	an illustrated tour of john boyegas journey to…	80	13	0.0000	0.000	1.000	0.000	1	0
57	serious question has any other star wars chara…	408	71	-0.4570	0.056	0.944	0.000	1	0
59	now this is the episode ix that everyone would…	58	11	0.5994	0.000	0.719	0.281	1	0
73	lets be honest the reason this community was m…	184	35	0.5466	0.078	0.761	0.162	1	0
74	watching a node video during their soap box de…	159	30	0.0000	0.000	1.000	0.000	1	0

Conclusion

- Using advanced classification techniques, we can improve classification of highly similar subreddit posts from a baseline accuracy of 50.1% to close to 80%.
- Adding VADER sentiment scores to TFIDF vectorized text greatly improves model performance in this case
- Further fine tuning, a larger dataset, or a dataset more specific to the problem - such as posts that have been removed in the past, may continue to improve results

May the force
be with you!

Thank you!