Subreddit Post Classification - Don't r/AskCulinary if your goal is to r/EatCheapAndHealthy

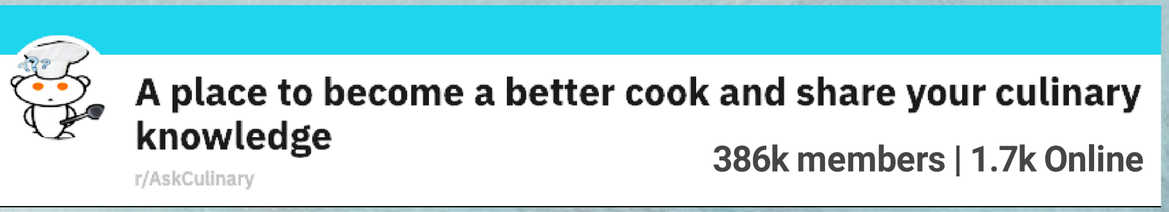
www.reddit.com

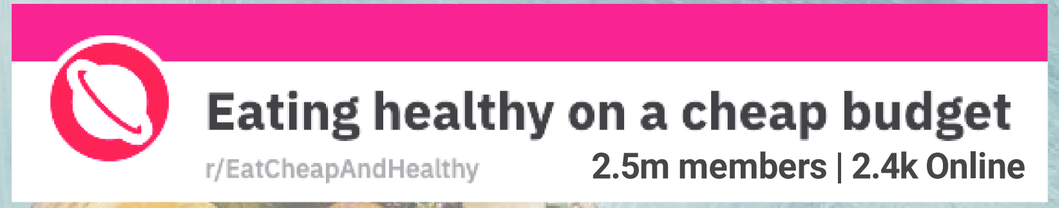
Reddit: the front page of internet

Reddit is a network of communities based on people's interests. Find communities you're interested in, and become part of an online community!

10 Reddit Statistic 2020

#.	Description
1	there are more than 430 million monthly active Reddit users worldwide (Reddit, 2019).
2.	it is currently the sixth most used social networking mobile app in the United States (Statista, 2019).
3.	over half (50.78 percent) of its desktop traffic originates from the US,
4.	Reddit users spend 10 minutes and 23 seconds on the site per visit (Similarweb, 2020).
5.	the site is the most popular among users in the 25 to 29 age group (Marketing Charts, 2019)
6.	there are more than 2,2 million subreddits (Reddit Metrics, 2020).
7.	there are currently more than 130,000 active communities on the platform (Reddit, 2019).
8.	weekends and Mondays are the best time to post on Reddit (X-Cart, 2020).
9.	titles of between 60 and 80 characters perform the best and receive the highest amount of upvotes.
10.	Reddit users are now watching a whopping 1.4 billion native videos on the platform





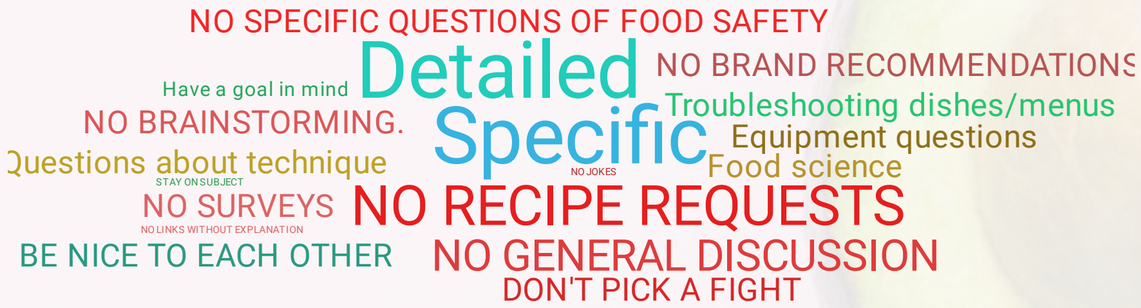
Data Gathering approach

- Data is clollected through Reddit API
- Data was collected through three different dates
- EDA was peformed against data from first request
- Final analysis was performed against datasets from three dates concated together.

Can we catch an NG post before it's published?

- Ask Culinary have a very specific guideline to its member for what to post and what not to post
- Show a pop-up window to the user if their posts violate the rule
- Asking users to look into other subreddit or give them search result based on their post content

r/AskCulinary: Dos and Don'ts



Dataset overview

Questions & Hypothesis

- Is the length of titles from two subreddits different?
- Hypothesis: since askculinary asks its member to be specific with detail, the title of its post should be longer.

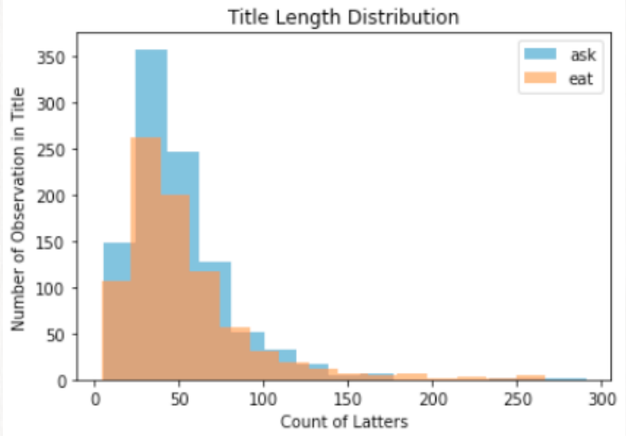
- What are words used or more frequently used in one subreddit than the other?
- Even though askculinary asks its members not to request recipes, I assume we will still see the word recipe and recipes often
- People from time to time need to provide recipe to build the context of their questions.

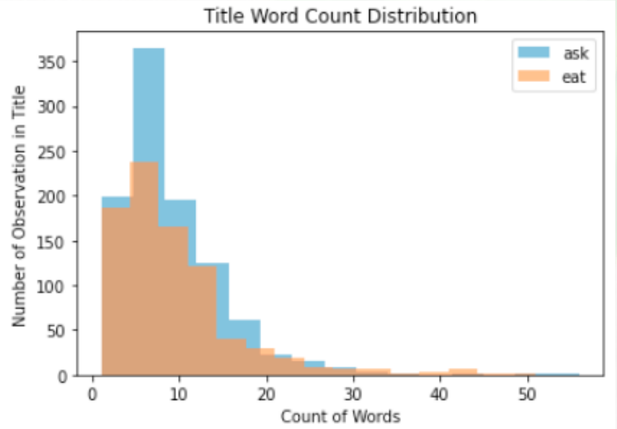
Since askculinary is more technical, skills, and equipment centric, does the title reflect this orientation?

- Is the post from either subreddit emotional?
- Because of the orientation of both subreddits, my hypothesis is:
- "No, the post shouldn't be emotional or with significant sentiment difference."

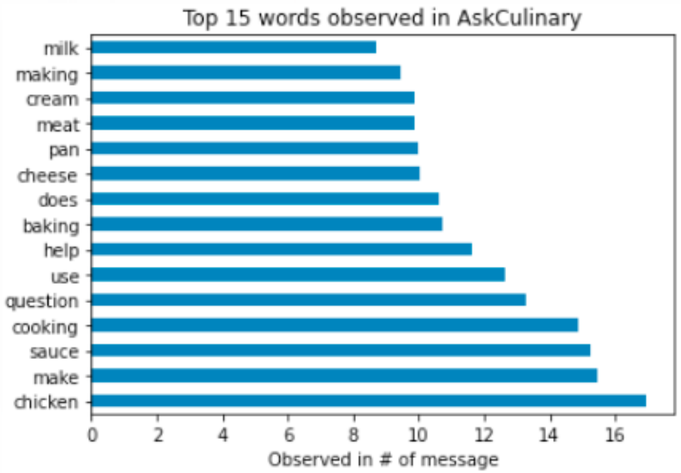
Clean-up and EDA

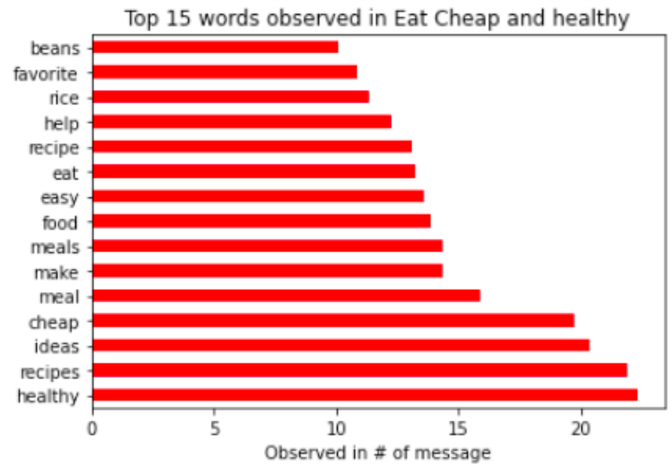
- Not much clean-up: no empty titles,
- Not using body: url, images, missing data, see title, lengthy
- Difference between post length?
- Difference between title length and word count?
- What are the frequently observed words?
- What are the Bigran and Trigram



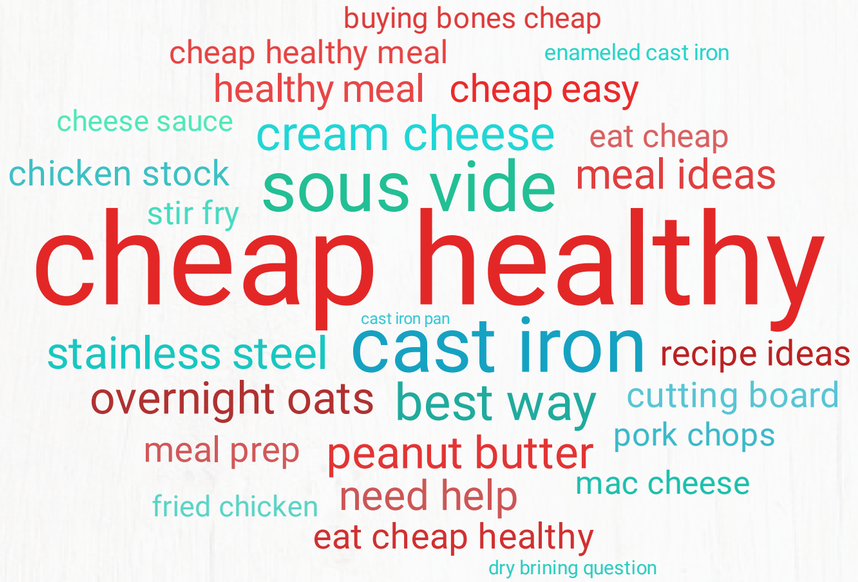


Top 15 Frequently Observed Words





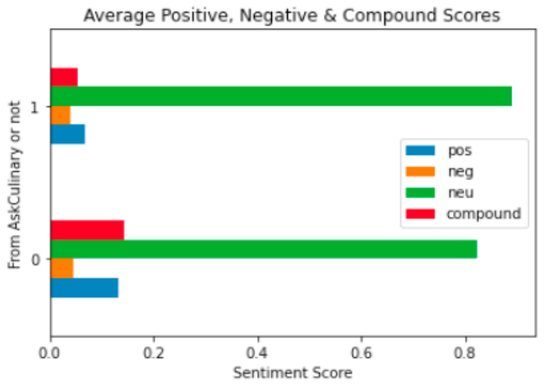
Bigram & Trigram



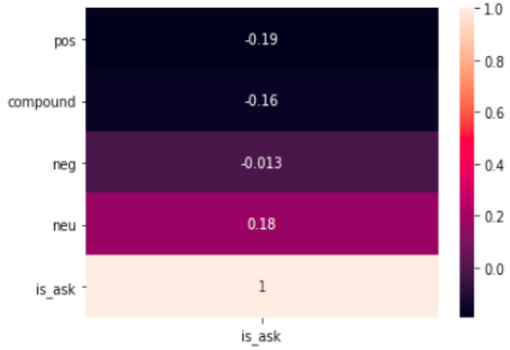
Are those posts emotional ?

	neg	neu	pos	compound
count	1834	1834	1834	1834
mean	0.042834	0.860178	0.096992	0.093886
std	0.114114	0.193101	0.168195	0.286716

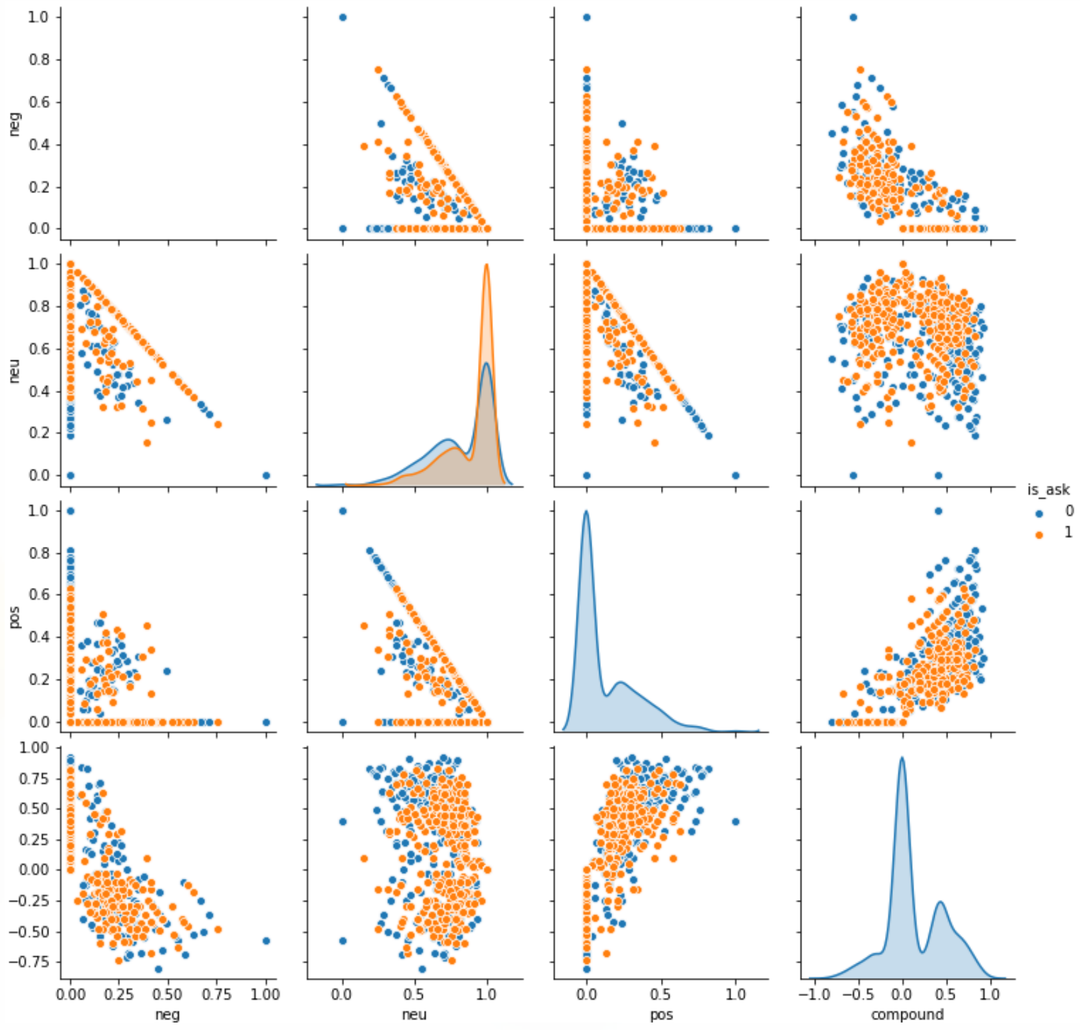
Average positivity, negativity and compound scores



Observation of Correlation?



Observation of Sentiment score and posts



Modeling & Evaluation

Various Calssification models:

- Pre-process with TfidfVectorizer
- Pipeline & GridSearch

Classification Model & Accuracy Score

- **Baseline score: 0.5071**

Pre-processing & Parameters	Model	Train Score	Test Score	fits
TfidfVectorizer Example of Best params: {'rf__max_depth': None, 'rf__max_features': 'auto', 'rf__n_estimators': 100, 'tvec__max_features': 2500, 'tvec__ngram_range': (1, 1), 'tvec__stop_words': 'english'} vote_params: 'ada__n_estimators': [50,75], 'gb__n_estimators': [100,125], 'tree__max_depth': [None, 5]	Logistic	0.917	0.7995	62
	KNN	0.864	0.7516	62
	Naive Bayes	0.9381	0.8235	91
	Random Forest	0.9994	0.8548	
	Bagging	0.9989	0.8201	
	AdaBoost	0.9994	0.8296	
	GradientBoost	0.9421	0.7949	
	Voting	0.88	0.7728	
	Linear SVC		0.8580	
	Kernel SVC	0.9936	0.8643	
	Kernel SVC C3	0.9989	0.8690	

Confusion Matrix & Evaluation

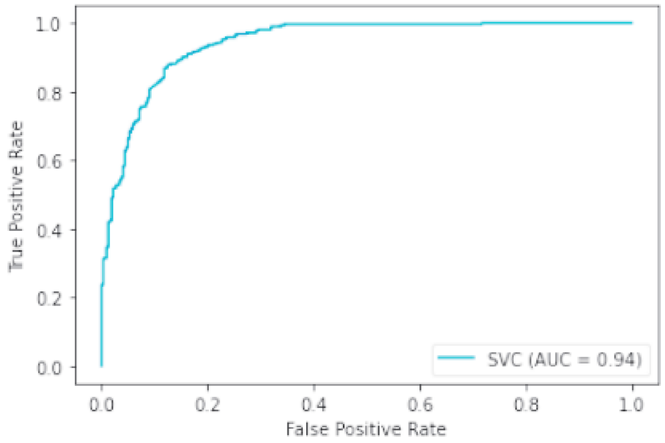
0.869	Predicted 0	Predicted 1
Actual 0	270	52
Actual 1	31	281

0.8643	Predicted 0	Predicted 1
Actual 0	261	61
Actual 1	25	287

Metrics	Score
Accuracy	0.8690
Sensitivity	0.90
Specificity	0.8385
Percision	0.8438

Metrics	Score
Accuracy	0.8643
Sensitivity	0.9198
Specificity	0.8105
Percision	0.8247

ROC

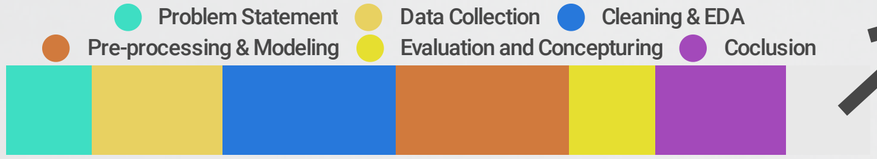


Considering how similar the two subreddits are, whether the model can diffirentiate posts from one another was a major concern.
The model can do better, but it's not too bad the way it is.

Are we ready to show pop-up messages to users based on the content of their posts?

Process:

90% Done



What's next?

- Can we get more data to improve the performance?
- Maybe adapt other tokenize approach to fine tune the data?
- Should we focus on improving accuracy, sensitivity, or specificity?
- Can we have the conersation with the community members and managers to understand the issue from their perspectives?

Finding from the research

- Posts from AskCulinary are not necessary more lengthy.
- Title length and word counts are actually very similar.
- After stop words were removed, we can see the vocabulary used are very different and it does reflect the theme of the community.
- More techincal terms, skills and equipments related word observed in AskCulinary posts.
- Posts are netural from both communities.

Improve Sensitivity

Community manager might not like it

More False Positive posts into discussion board, diffuse the purpose of this project.

Improve Specificity

Members might not like it

False negative triggers pop-ups asking members to go somewhere else when it's legit.