P3: Subreddit Classifier

How to identify which subreddit the post came from?

Table of contents

01

Scenario

02

Problem Statement

03

Subreddits

04

EDA

05

Models

06

Conclusion & Business Recommendation

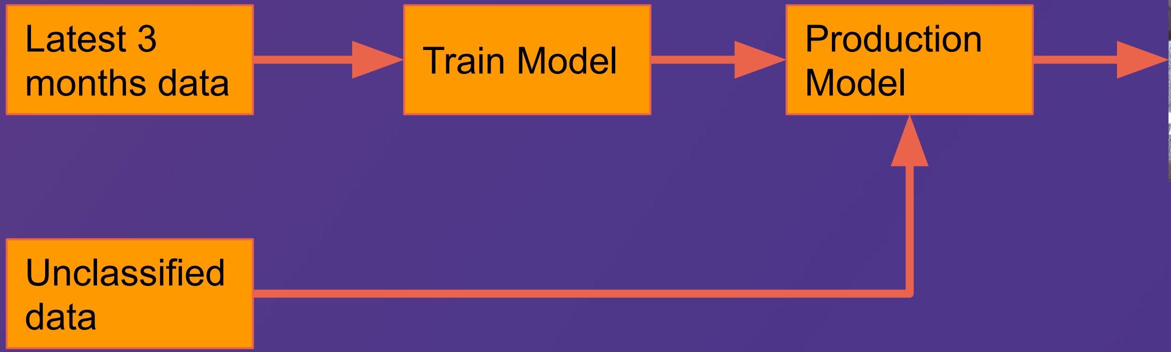
Scenario

Due to multiple cyber attacks recently, many reddit posts which were 3 months and older were taken offline as their data were held hostage and copies were deleted. As the management team had decided to not give in to the demands of the hackers, the original data were not recovered. Luckily, the IT team has managed to recover the data partially with informations such as the title, selftext and etc. Unfortunately, those data recovered did not have the subreddit name, url and links that would provide identifications to those posts.

Problem Statement

Being a data scientist in Reddit, you made a suggestion to your manager that perhaps the subreddit name could be inferred/predicted from the remaining information recovered through modeling. As a proof of concept, you have been tasked by your manager to:

- Use the latest 3 months of data, complete with subreddit name, from 2 random subreddits
- Determine the feasibility and accuracy of the suggested approach in identifying the reddit posts
    - Target goal is to achieve at least 90% accuracy



Latest 3 months data

Train Model

Production Model

Unclassified data

Subreddits

LinusTechTips

TrashTaste

Subreddits - LinusTechTips

- YouTube channel
- 14.4 millions subscribers
- Revolves around technology, computers, laptops, phones, reviews and funny but impractical proof of concepts
- Has multiple sister channels (e.g. TechLinked, Techquickie)

Linus

Anthony

Riley

Luke

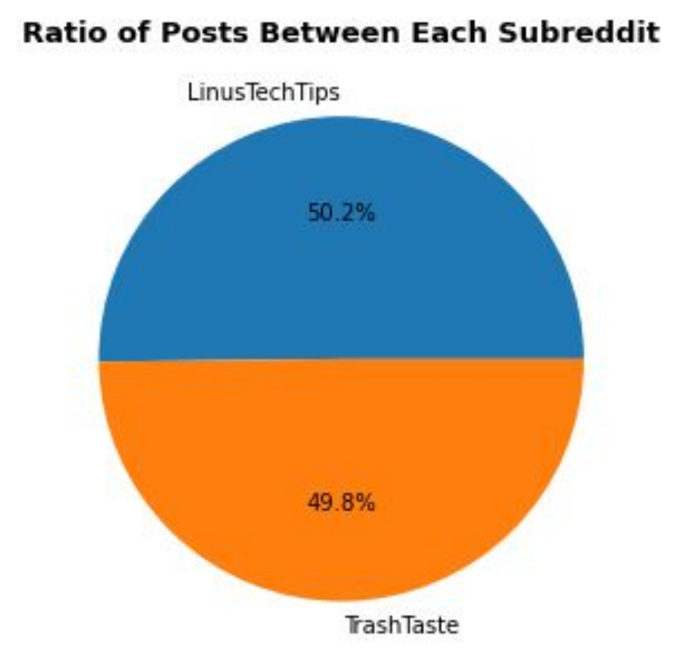
Subreddits - TrashTaste

- YouTube channel
- 1.22 millions subscribers
- Podcast with no specific genre
    - Hosts gives their take on different topics

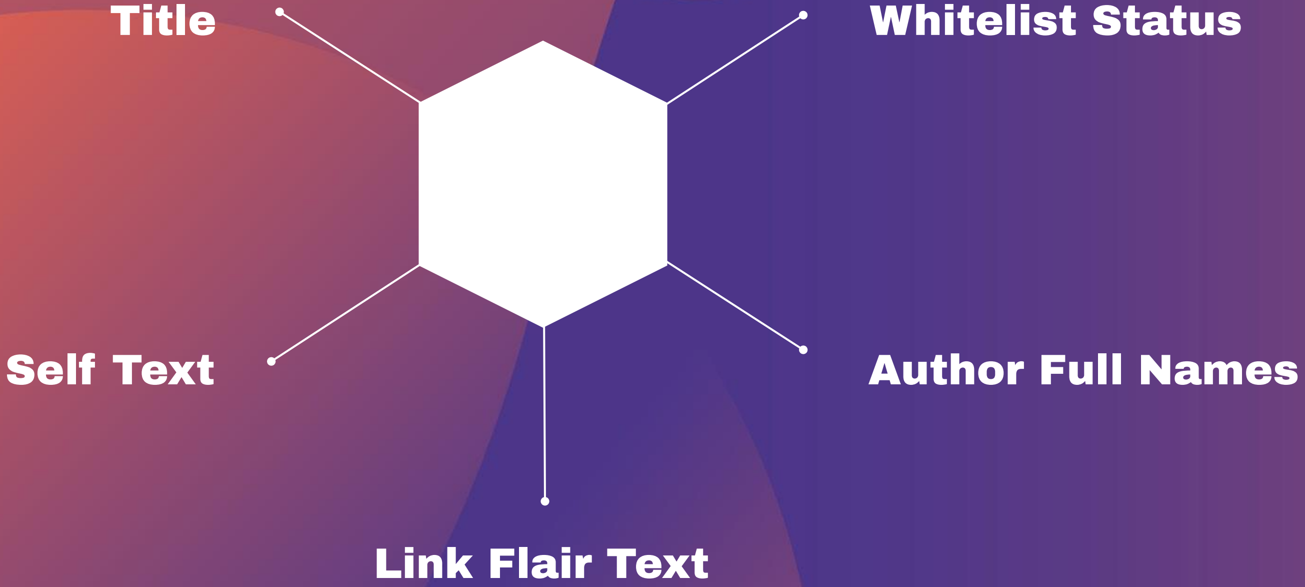
Joey Connor Garnt

Base Model

As the dataset is split roughly into half between the 2 subreddits, the base model would be to assume all posts as LinusTechTips and the base model accuracy would be 50.2%.



Selected Predictors



Title

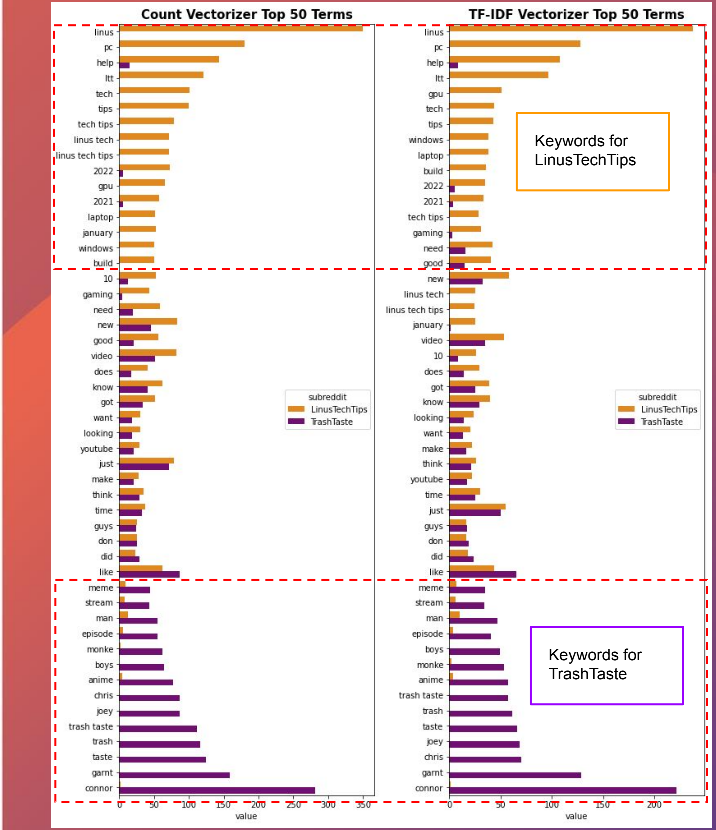
Whitelist Status

Self Text

Author Full Names

Link Flair Text

Useful Keywords in Title



Keywords for LinusTechTips

Keywords for TrashTaste

While there are many words that are common on both subreddits, there are some keywords that helps to identify which subreddit each post is coming from.

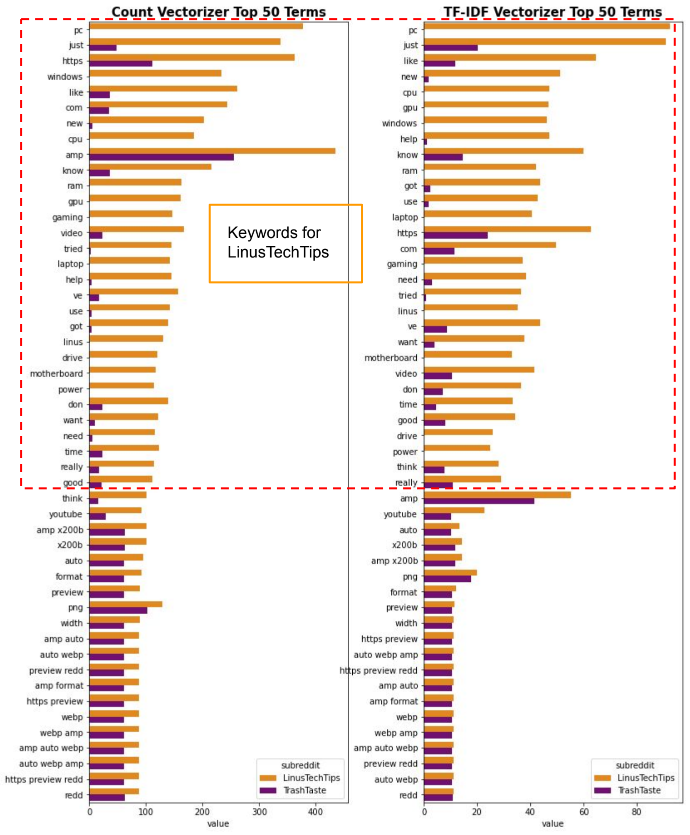
TrashTaste

- connor, garnt, chris, joey
  - Names of the hosts and guests
- boys
  - Nickname of the hosts

LinusTechTips

- linus
  - Name of the host/owner
- pc, gpu, laptop, build, gaming
  - Computer related terms

Useful Keywords in Self Text


Count Vectorizer Top 50 Terms

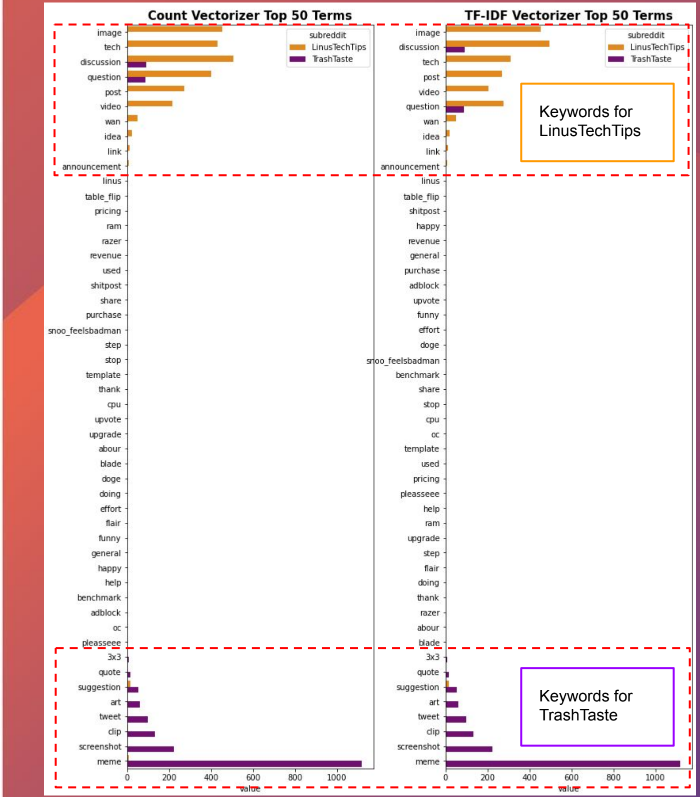
TF-IDF Vectorizer Top 50 Terms

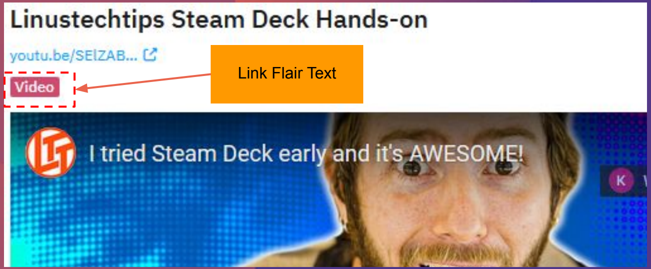
Keywords for LinusTechTips

Unfortunately in self text, there were not many distinct keywords that would allow post from TrashTaste to distinguish itself from LinusTechTips.

Keywords for LinusTechTips are still mainly computer related.

Useful Keywords in Link Flair Text





Linustechtips Steam Deck Hands-on

youtu.be/SElZAB...

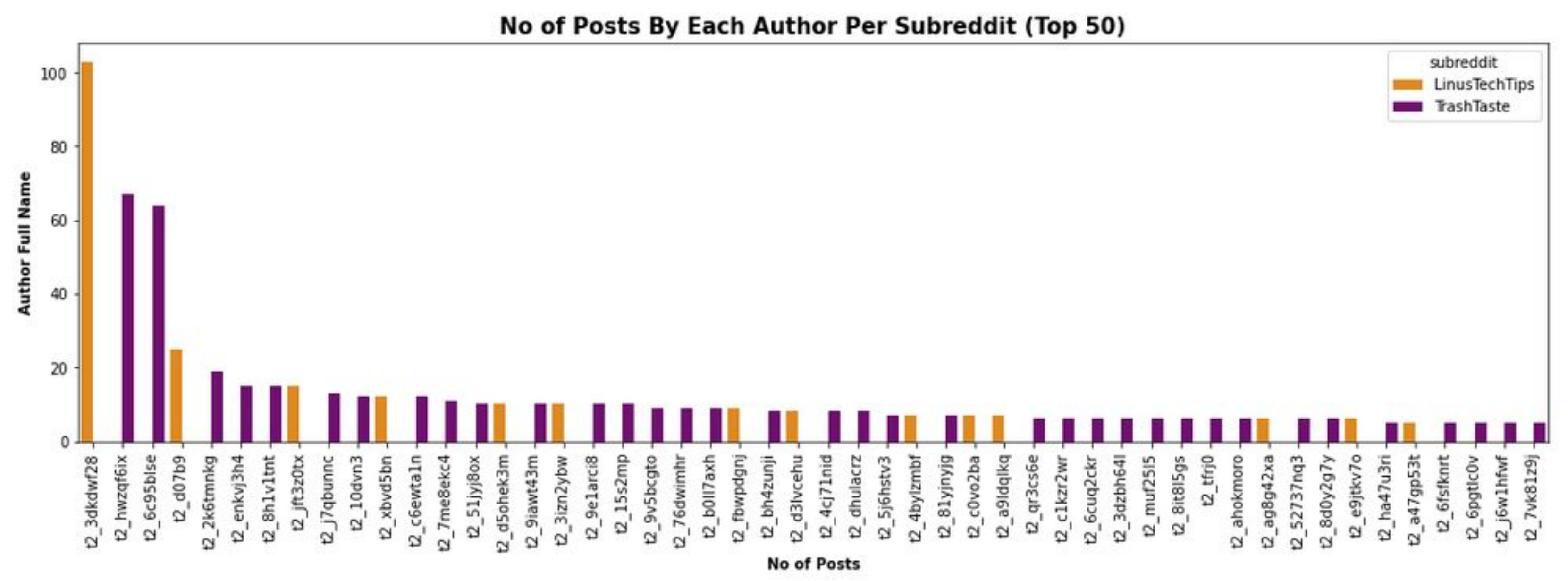
Video

Link Flair Text

I tried Steam Deck early and it's AWESOME!

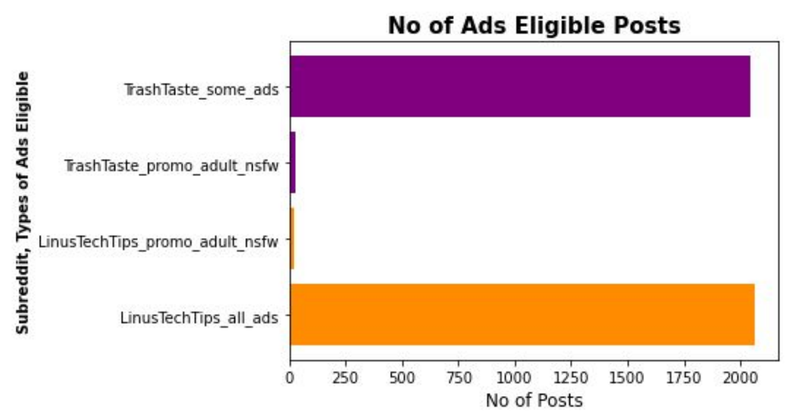
- There are not many distinct keywords for both subreddits
- But keyword “meme” appears very frequently as compared to other keywords
  - Might help to identify TrashTaste posts

Any Common Authors?



No overlap in authors between the 2 subreddits for the top 50 authors

Are Advertisement Allowed?

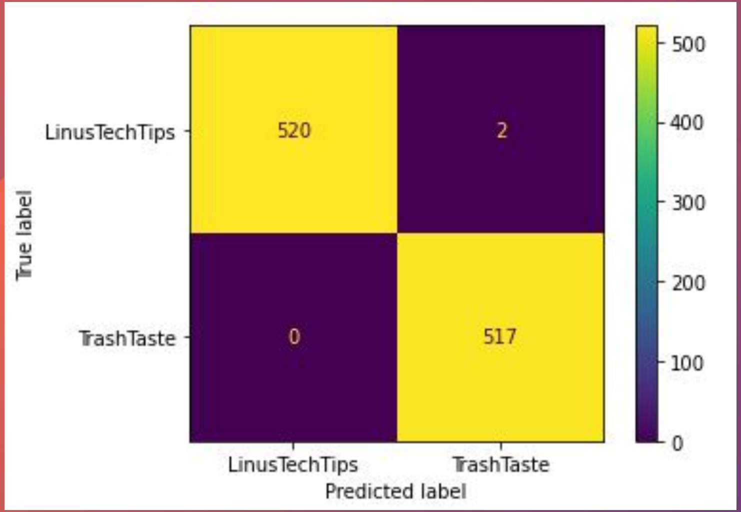


Very significant difference in the amount of ads that is allowed to be shown on the reddit posts:-

- TrashTaste - Only some ads are allowed
- LinusTechTips - All ads are allowed

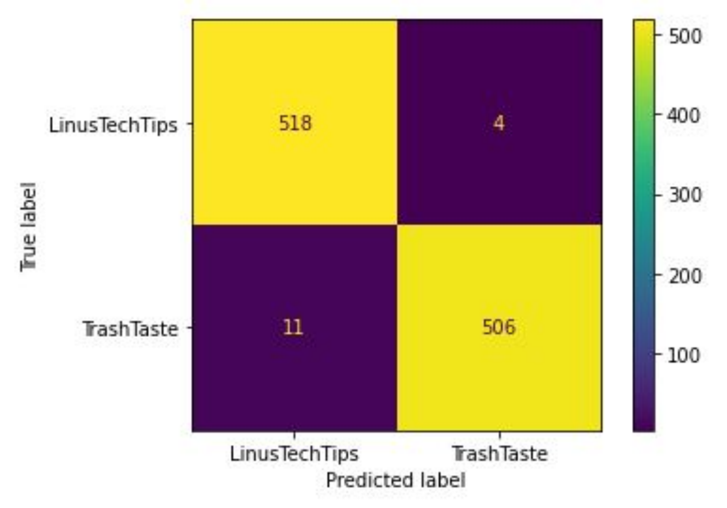
Models Performance

TF-IDF + Logistic



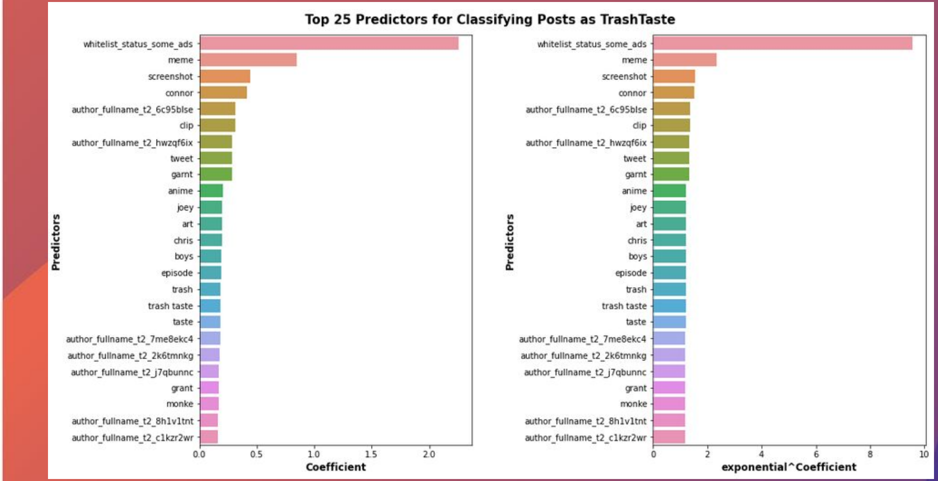
Accuracy: 99.8% >> 50.2% (Base model)
Sensitivity: 100%
Specificity: 99.6%

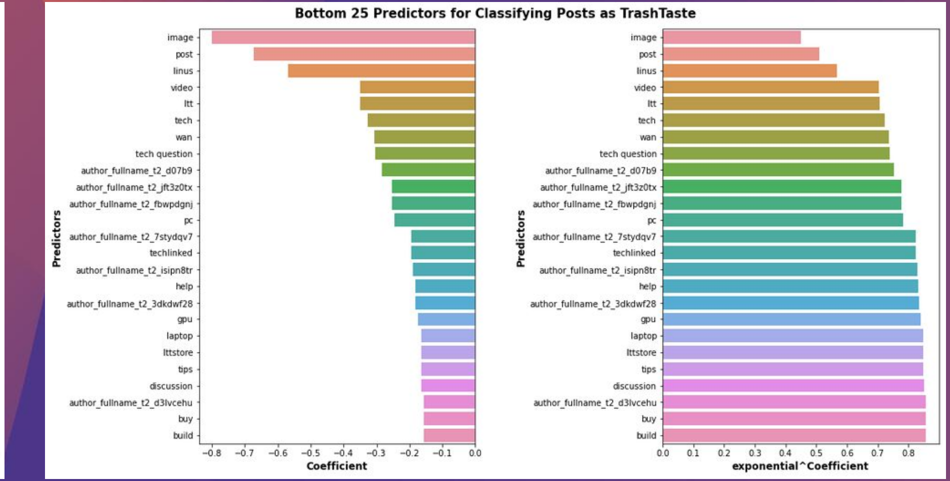
TF-IDF + Random Forest



Accuracy: 99.5% >> 50.2% (Base model)
Sensitivity: 97.8%
Specificity: 99.2%

Model 1 (TF-IDF + Logistic) Insights





Top predictors in increasing the odds of classifying posts as TrashTaste are ‘whitelist_status_some_ads’, ‘meme’, ‘screenshot’, ‘connor’.

Bottom predictors in decreasing the odds of classifying posts as TrashTaste are ‘image’, ‘post’, ‘linus’, ‘video’ and ‘ltt’.

Conclusion

Is it possible to use information from reddit posts to identify which subreddit they came from?

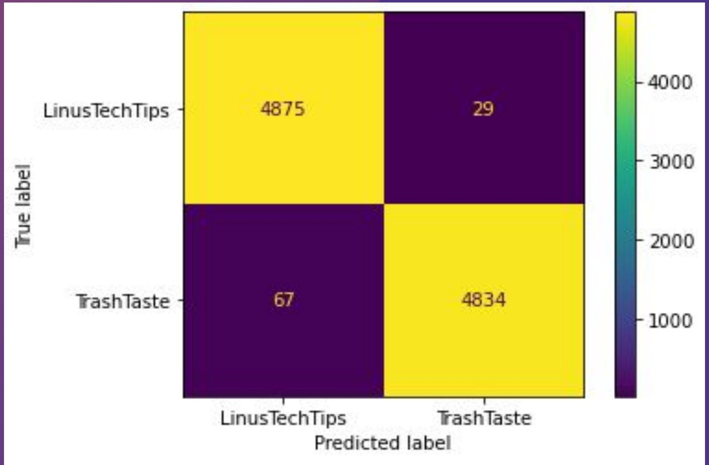
Yes!!!!

Accuracy is more than 90%!!!

Conclusion

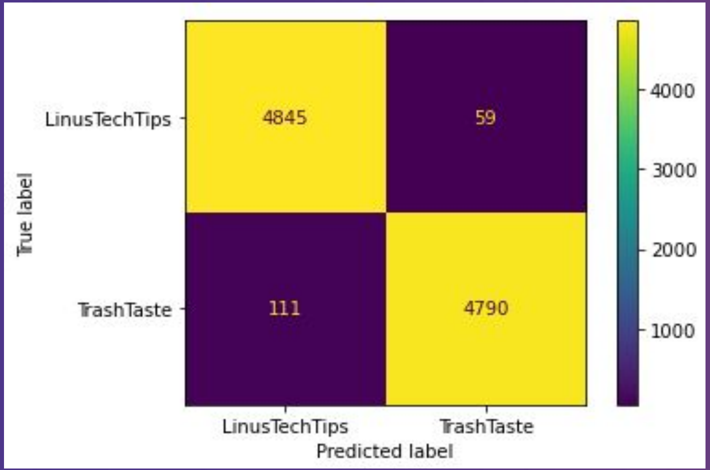
Accuracy on dataset from late 2020 is maintained at above 90%.

TF-IDF + Logistic



Accuracy: 99.0%
Sensitivity: 98.6%
Specificity: 99.4%

TF-IDF + Random Forest



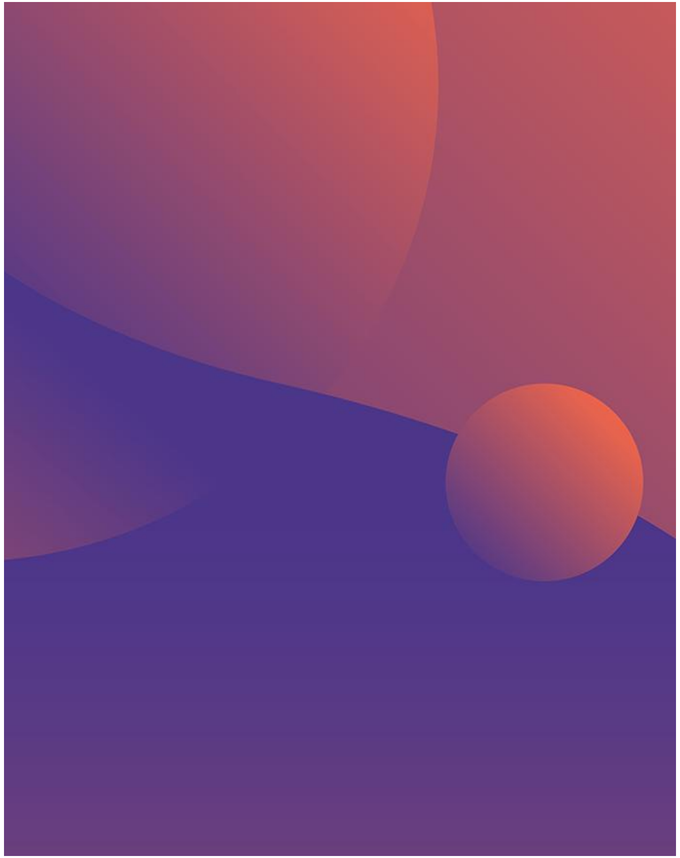
Accuracy: 98.2%
Sensitivity: 97.7%
Specificity: 98.7%

Business Recommendations

- Reddit could show a list of subreddits to the users which the posts would be suitable to be posted in
- With the logistic regression model, since there are coefficients that indicate what are the key terms for the subreddits, Reddit could list down the current trend or hot words in the subreddit

Q&A

Appendix

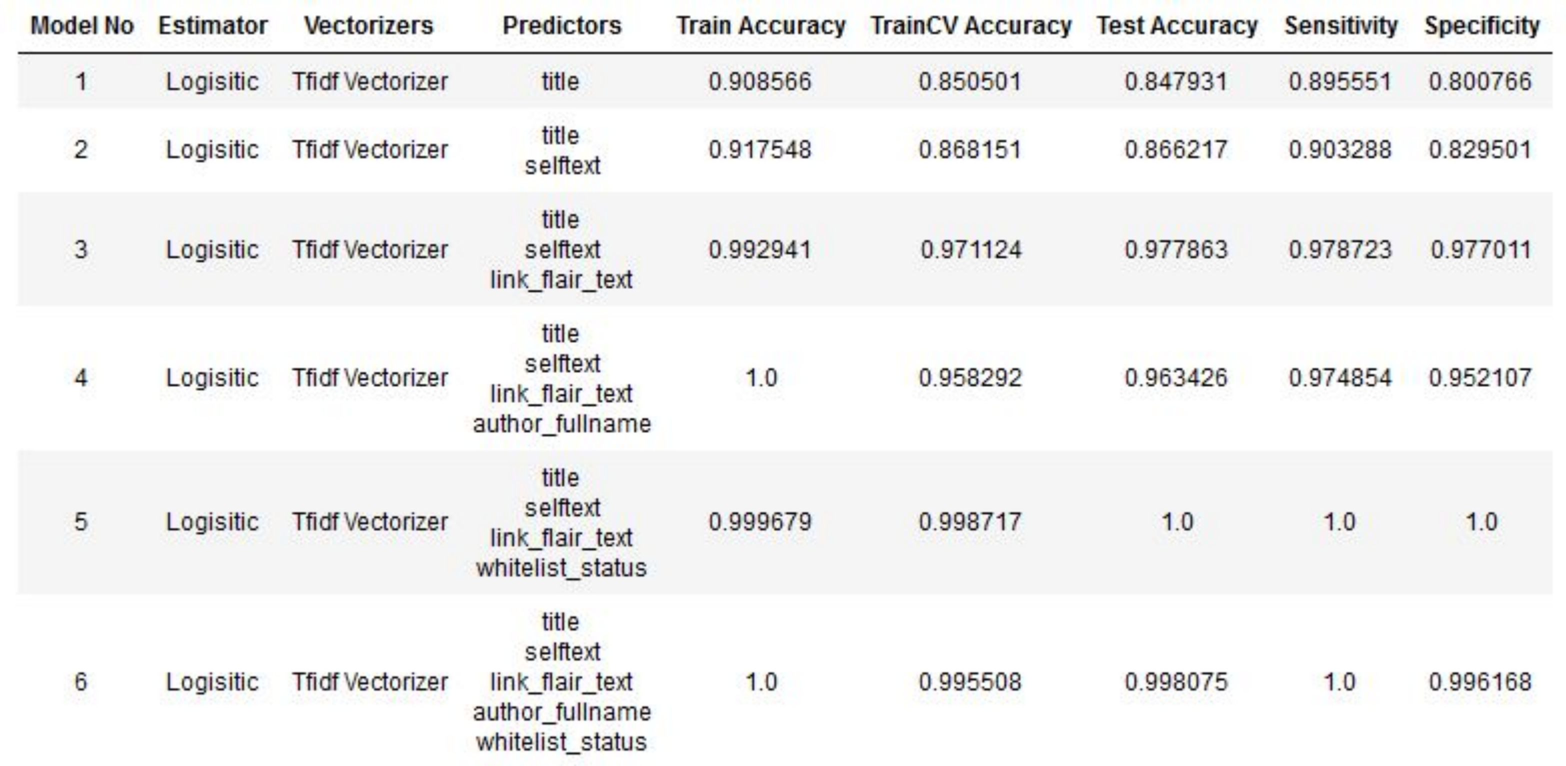


Potential Improvements

As the reason this current production model is showing very high accuracy is due to the significant difference between the whitelist_status of the 2 subreddit chosen, when training model for other subreddits, we could:-

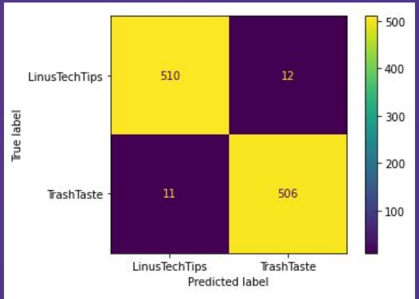
- Include comments from each subreddit post
- author_flair_richtext if the usage rate is high for other subreddits
- Increasing the amount of dataset used for training
- Checking if there are images in the post
    - Certain subreddits tend to have more image in post

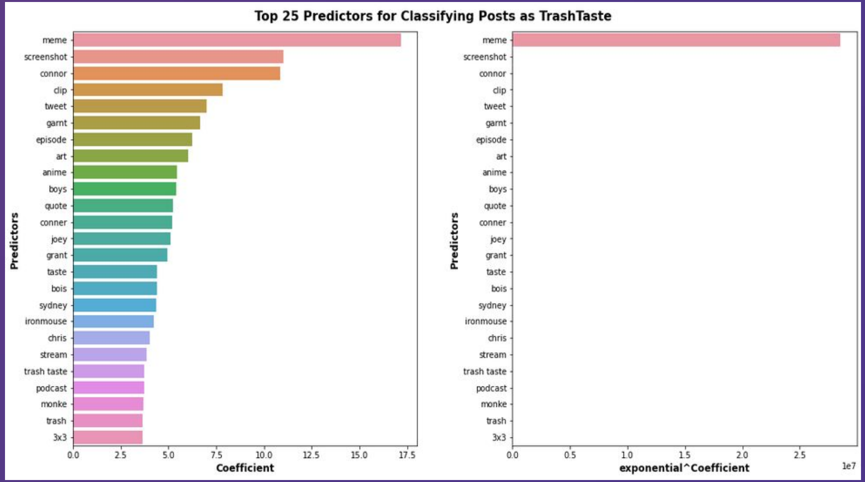
Accuracy for Other Combinations of Predictors (LR)

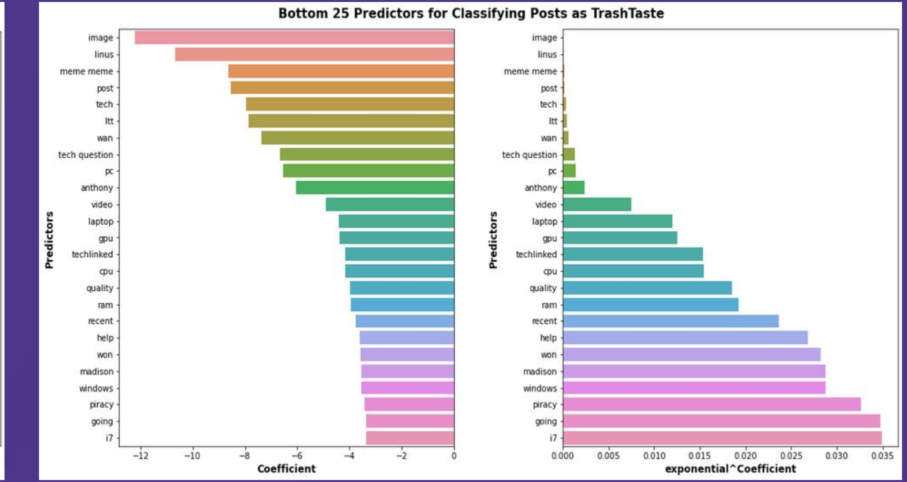


Model No	Estimator	Vectorizers	Predictors	Train Accuracy	TrainCV Accuracy	Test Accuracy	Sensitivity	Specificity
1	Logisitic	Tfidf Vectorizer	title	0.908566	0.850501	0.847931	0.895551	0.800766
2	Logisitic	Tfidf Vectorizer	title selftext	0.917548	0.868151	0.866217	0.903288	0.829501
3	Logisitic	Tfidf Vectorizer	title selftext link_flair_text	0.992941	0.971124	0.977863	0.978723	0.977011
4	Logisitic	Tfidf Vectorizer	title selftext link_flair_text author_fullname	1.0	0.958292	0.963426	0.974854	0.952107
5	Logisitic	Tfidf Vectorizer	title selftext link_flair_text whitelist_status	0.999679	0.998717	1.0	1.0	1.0
6	Logisitic	Tfidf Vectorizer	title selftext link_flair_text author_fullname whitelist_status	1.0	0.995508	0.998075	1.0	0.996168

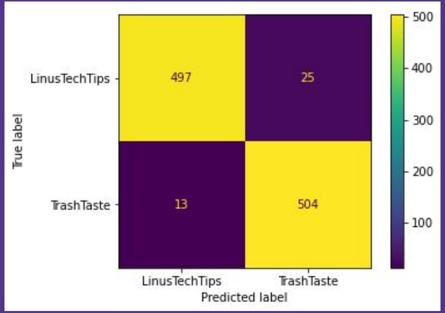
Model 3 Score and Coefficients (LR)

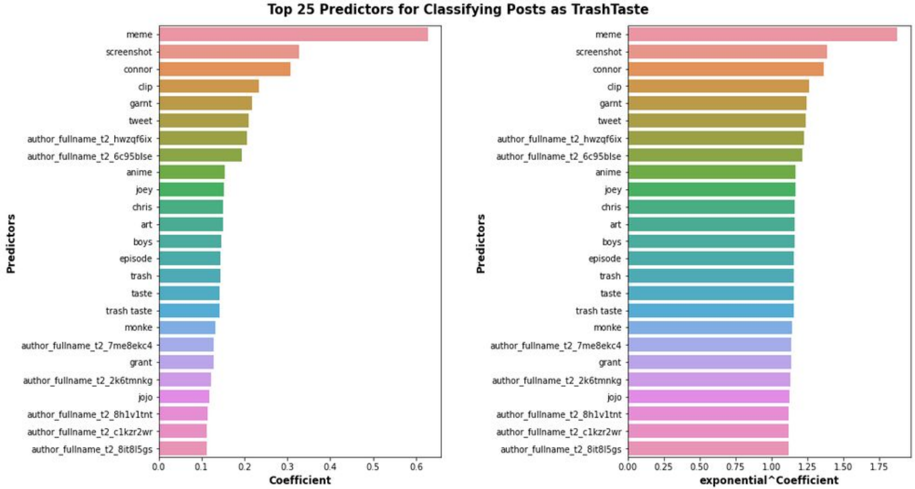


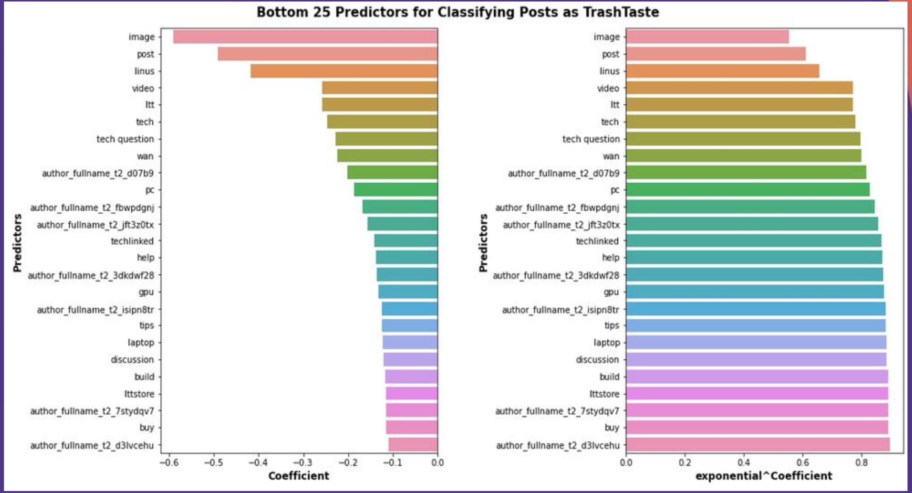

Top 25 Predictors for Classifying Posts as TrashTaste


Bottom 25 Predictors for Classifying Posts as TrashTaste

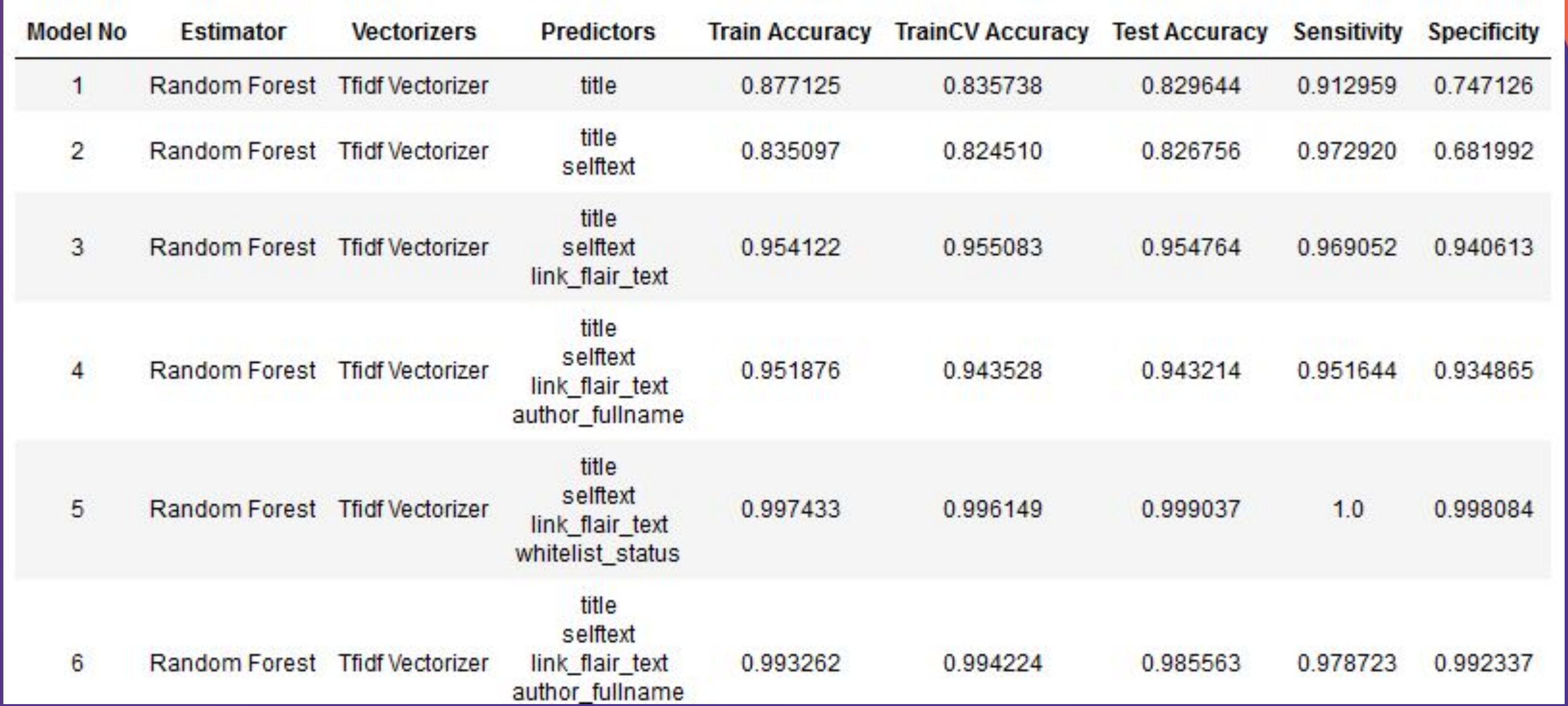
Model 4 Score and Coefficients (LR)




Top 25 Predictors for Classifying Posts as TrashTaste


Bottom 25 Predictors for Classifying Posts as TrashTaste

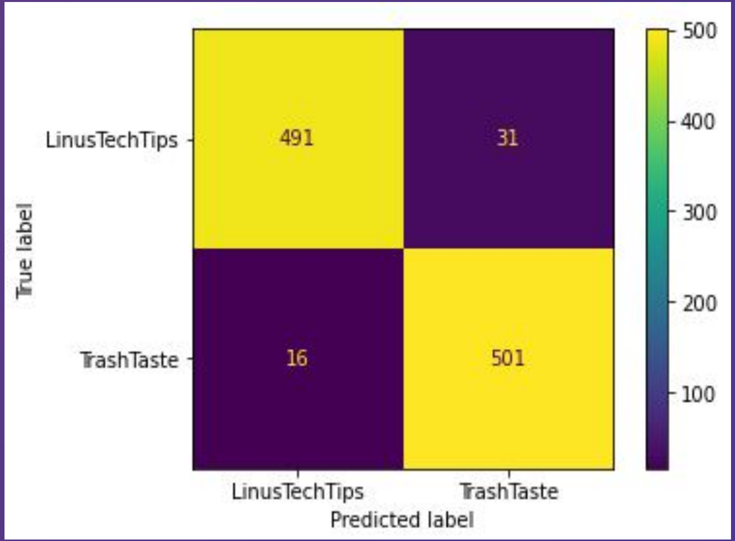
Accuracy for Other Combinations of Predictors (RF)



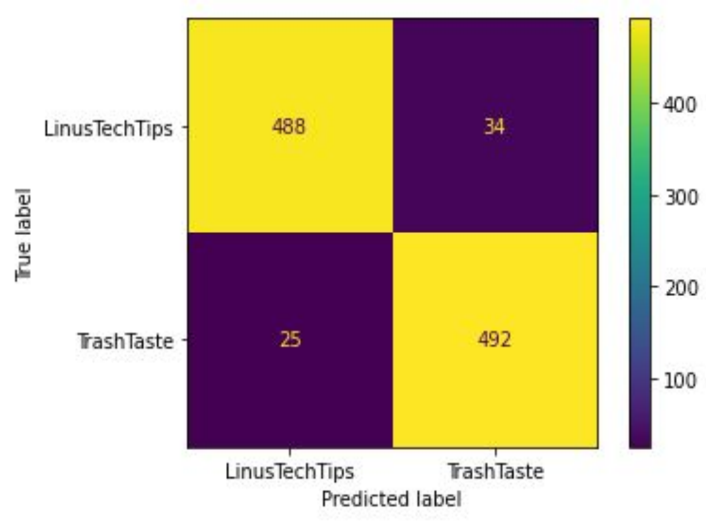
Model No	Estimator	Vectorizers	Predictors	Train Accuracy	TrainCV Accuracy	Test Accuracy	Sensitivity	Specificity
1	Random Forest	Tfidf Vectorizer	title	0.877125	0.835738	0.829644	0.912959	0.747126
2	Random Forest	Tfidf Vectorizer	title selftext	0.835097	0.824510	0.826756	0.972920	0.681992
3	Random Forest	Tfidf Vectorizer	title selftext link_flair_text	0.954122	0.955083	0.954764	0.969052	0.940613
4	Random Forest	Tfidf Vectorizer	title selftext link_flair_text author_fullname	0.951876	0.943528	0.943214	0.951644	0.934865
5	Random Forest	Tfidf Vectorizer	title selftext link_flair_text whitelist_status	0.997433	0.996149	0.999037	1.0	0.998084
6	Random Forest	Tfidf Vectorizer	title selftext link_flair_text author_fullname	0.993262	0.994224	0.985563	0.978723	0.992337

Model 3 and 4 Score (RF)

Model 3



Model 4


LinusTechTips
TrashTaste
True label
LinusTechTips
TrashTaste
Predicted label