

# Subreddit Classification

Ryan Pedersen  
DSI - 523



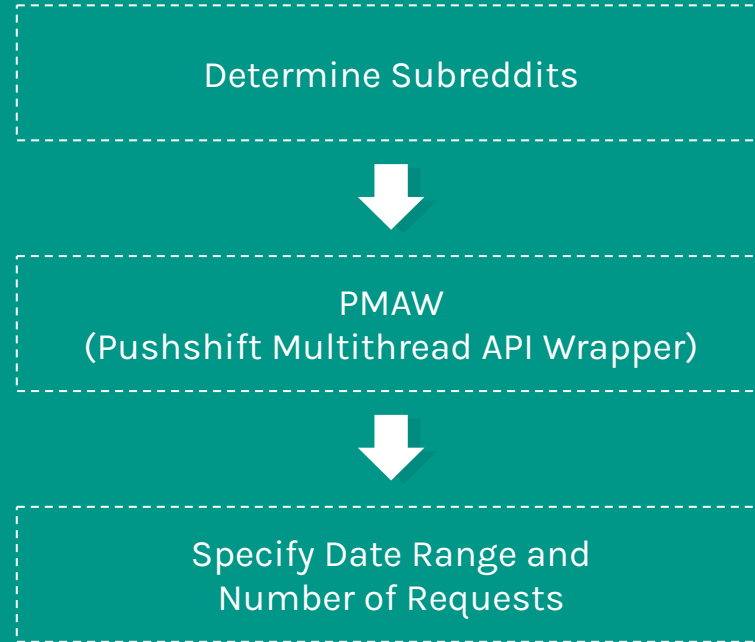
## Why?

1. Which subreddit to post on?
2. Where did a post come from?

# Data Collection



## Data Collection Process



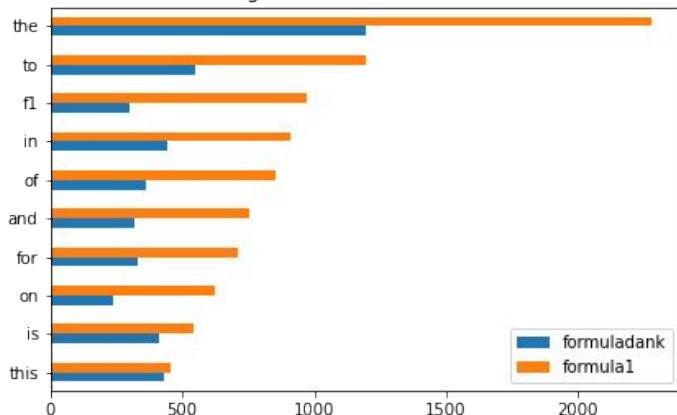
# Data Collection

- Subreddits chosen: r/formula1 and r/formuladank
  - Formuladank is Formula 1 memes.
- 10,000 comments per subreddit
- 5,000 posts per subreddit
- Date Range: 1/1/2021 thru 6/24/2022

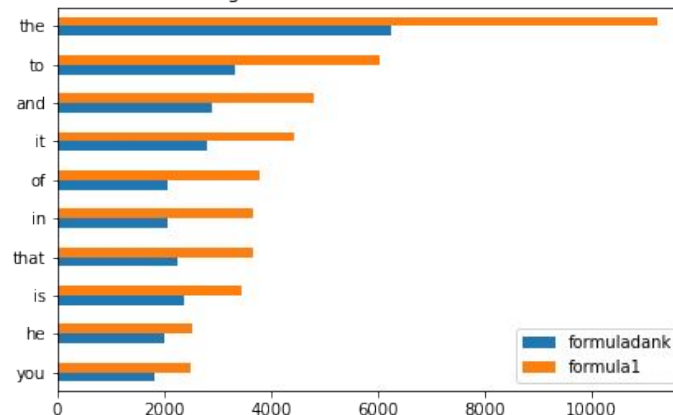
# Exploratory Data Analysis

# WORD COUNTS

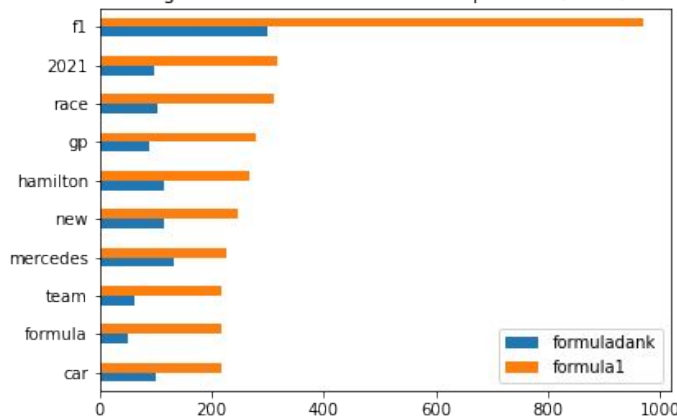
Highest word counts (Titles)



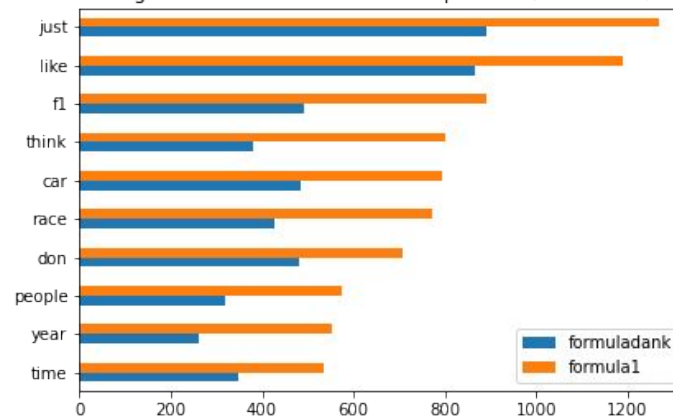
Highest word counts (Comments)



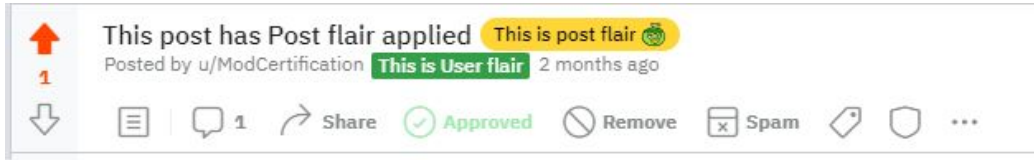
Highest Word Counts Without Stop Words (Titles)



Highest Word Counts Without Stop Words (Comments)



# Flairs



- 3135 flair-less users in all comments
  - 3088 in formula1
- 2413 flair-less users in all posts
  - 2205 in formula1

Formula1 flair: Max Verstappen

Formuladank flair: Don't know f1 but memes are kinda funny



# Modeling

# Roadmap

Build a simple model  
with no  
hyperparameters



Iterate on grid search  
parameters by adding  
more



Using the models that  
did well, grid search for  
hyperparameters



Should features be  
added or removed?  
Return to start



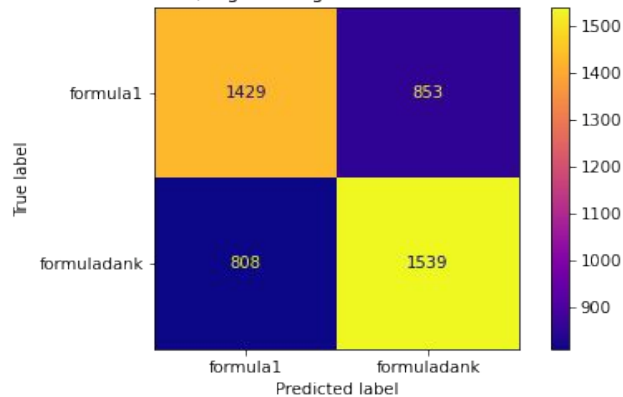
# Model Parameters and Accuracy

	TF-IDF & Logistic Regression (Comments)	Countvectorizer & Logistic Regression (Comments + Flair)	TF-IDF & Random Forest (Titles)	Countvectorizer & Logistic Regression (Titles + Flair)	Countvectorizer & Logistic Regression (Only flairs from comments)	Countvectorizer & Logistic Regression (Only flairs from titles)
Feature Extraction Parameters	Max df: 0.3 Max Features: 10,000	Defaults	Max df: 0.25 Max Features: 8,000	Defaults	Defaults	Defaults
Classifier Parameters	C: 0.75	Defaults	Num of estimators: 200	Defaults	Defaults	Defaults
Training Accuracy	75.5%	99.7%	99.8%	98.4%	99.5%	97.8%
Testing Accuracy	64.1%	99.3%	77.2%	97.1%	99.5%	97.3%

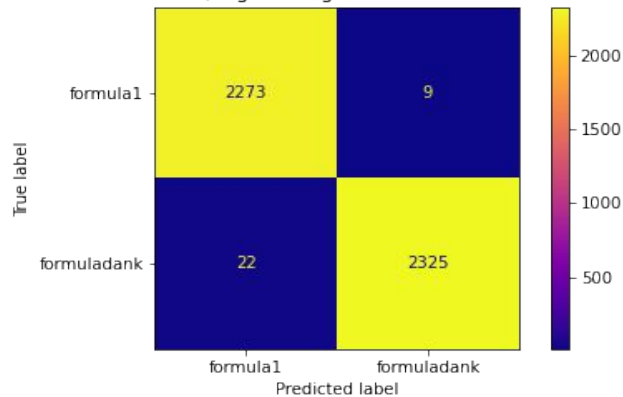
Baseline Accuracy	50%
-------------------	-----

# Confusion Matrices for Models

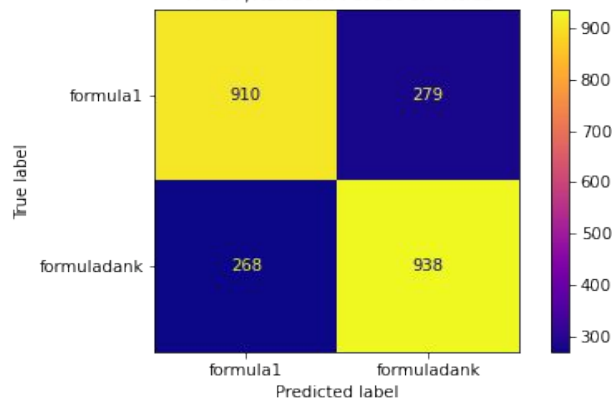
TF-IDF/Logistic Regression on Comments



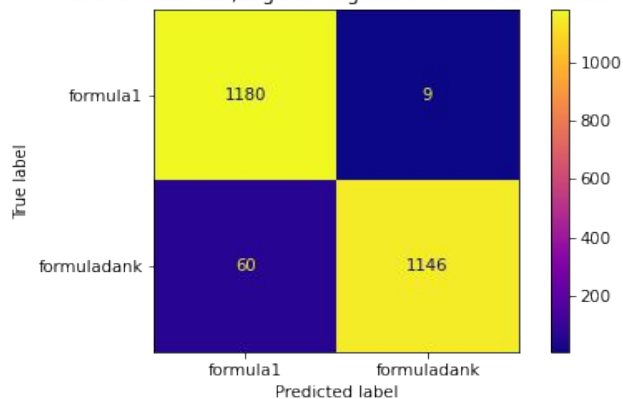
Count Vectorizer/Logistic Regression on Comments + Flairs



TF-IDF/Random Forest on Titles



Count Vectorizer/Logistic Regression on Titles + Flairs





## Next Steps

- ❑ Dive deeper into grid searching with flairs as a feature
- ❑ Model other similar subreddit pairs
- ❑ Find other interesting features to use



# Conclusions

- ❑ Building a model for two similar subreddits is very difficult
- ❑ Adding extra features can be extremely useful
- ❑ Complex models do not always outperform simple ones.

The simple  
model was  
better



The simple  
model was  
better





# THANKS!

## Any questions?

Special thanks to all the people who made and released these awesome resources for free:

- ▶ Presentation template by [SlidesCarnival](#)
- ▶ Photographs by [Death to the Stock Photo](#) ([license](#))