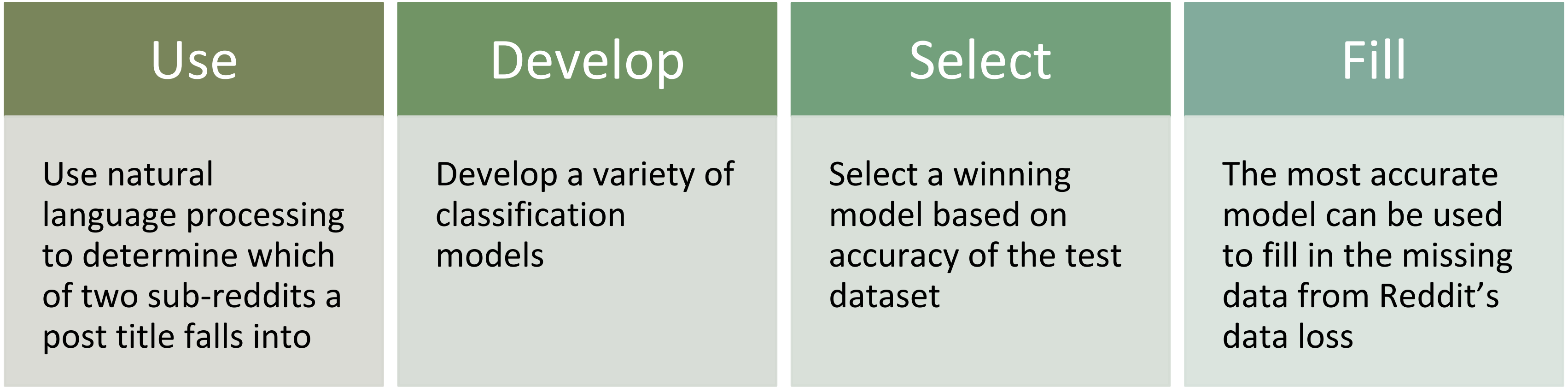
Up the Creek Without a Paddle:

Using Predictive Modeling to Address Reddit Data Loss

Problem Statement



Use

Use natural language processing to determine which of two sub-reddits a post title falls into

Develop

Develop a variety of classification models

Select

Select a winning model based on accuracy of the test dataset

Fill

The most accurate model can be used to fill in the missing data from Reddit’s data loss

Executive Summary

Collected post title and sub-reddit name using Reddit’s API

Dataset contains equal number of canoe and table tennis posts (balanced classes)

- Baseline accuracy is .50

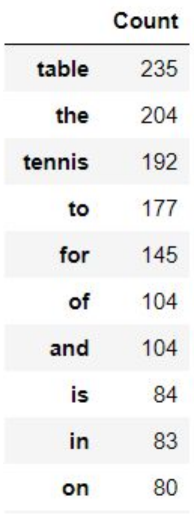
Ran multiple classification models using both CVEC and TF-IDF

All models have accuracy of at least .87 on the test dataset

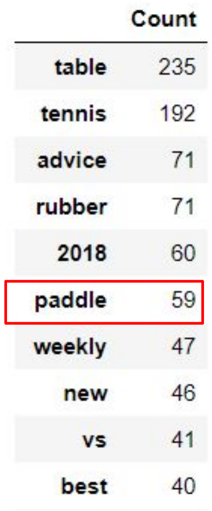
- Highest performer was logistic regression using TF-IDF (accuracy .93)

Documents contain many common words

Table Tennis top words



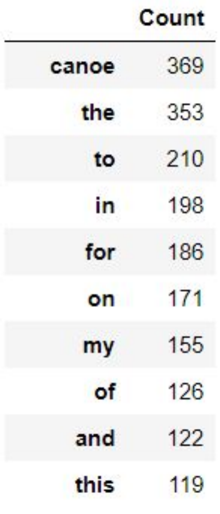
	Count
table	235
the	204
tennis	192
to	177
for	145
of	104
and	104
is	84
in	83
on	80



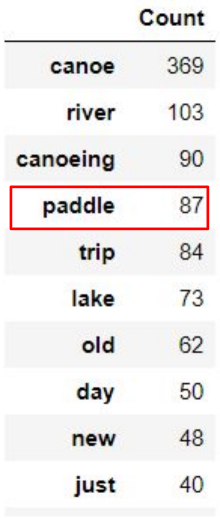
	Count
table	235
tennis	192
advice	71
rubber	71
2018	60
paddle	59
weekly	47
new	46
vs	41
best	40

Stop-words □ removed

Canoe top words



	Count
canoe	369
the	353
to	210
in	198
for	186
on	171
my	155
of	126
and	122
this	119



	Count
canoe	369
river	103
canoeing	90
paddle	87
trip	84
lake	73
old	62
day	50
new	48
just	40

Stop-words □ removed

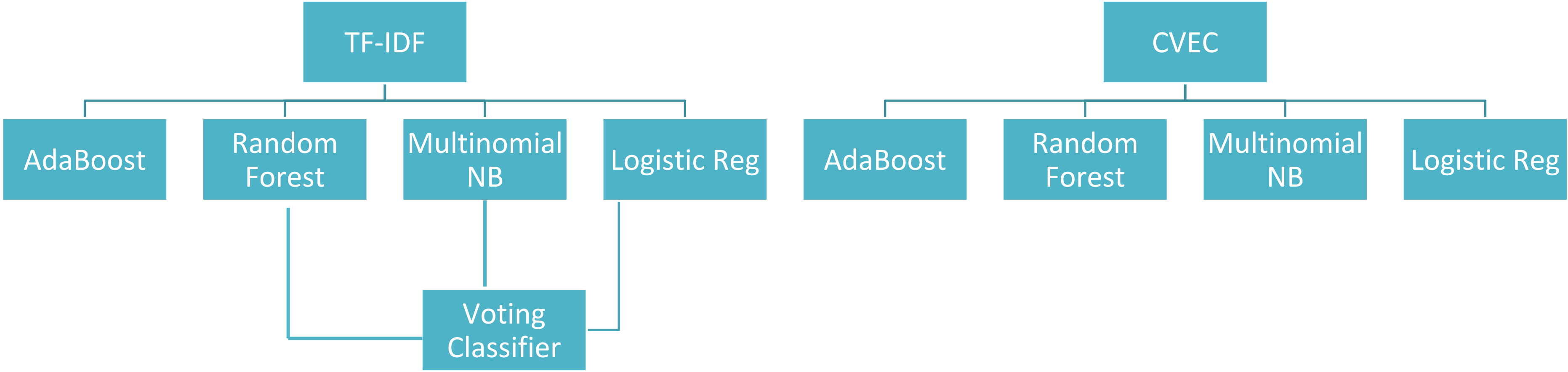
Using CVEC to explore the words included in the documents

Many are common stop-words, can be removed by CVEC

“Paddle” is a common word across both sub-reddits

Used both CVEC and TF-IDF to run models

Ran multiple models to find best performer



TF-IDF

AdaBoost

Random Forest

Multinomial NB

Logistic Reg

Voting Classifier

CVEC

AdaBoost

Random Forest

Multinomial NB

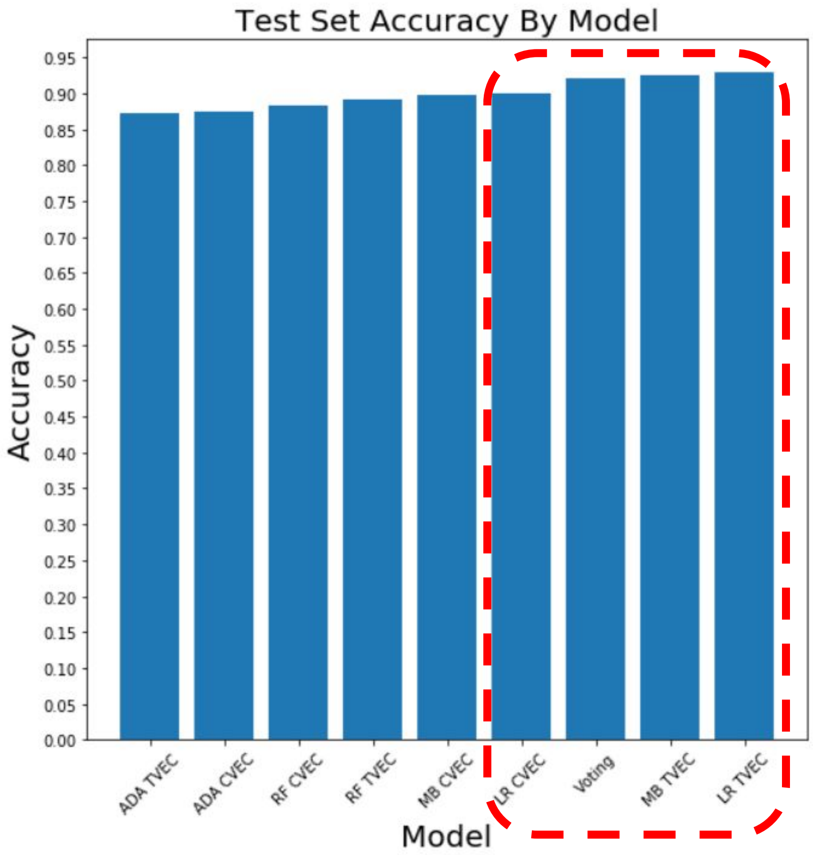
Logistic Reg

Four models have accuracy of .90 +

Logistic regression using TF-IDF (accuracy = .93), Multinomial NB + TF-IDF (.93), Voting Classifier + TF-IDF (.92), and Logistic regression using CVEC (.90) were the top performing models

AdaBoost, using both TF-IDF (.87) and CVEC (.87), were the lowest scoring models overall

All models outperform the baseline accuracy (.50)



Results are strong, but additional research can be performed

Overall, the models perform well at classifying the post titles

There are multiple top-performing models to choose from and implement

Experimentation with Hashing-Vectorizer and additional Voting Classifiers may yield strong results

Questions?