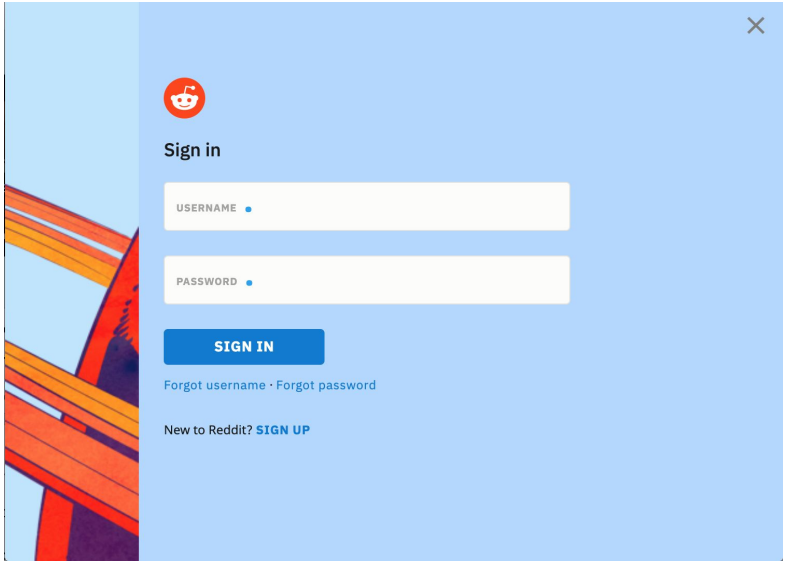
Tripping Down Classification Road

How to Distinguish Between The Absurd
Vs. The Absurd But True

Problem Statement

To assemble posts from two chosen subreddits, which in turn serves as feedstock to train a classification model that has the capability to bucketize according to a posting's origin.



But Which Two ??

GOLDILOCKS TASK:

- Avoid two topics that are completely unrelated, making the classification too easy.
- Avoid two topics that are the same, making the classification too hard.
- Take two that have substantial enough overlap to make classification challenging but doable.

A FEW EXAMPLES OF MISFIRES ON MY PART:

Superclass and a subclass --- It’s winner take all, so subclass doesn’t exist or too sparse

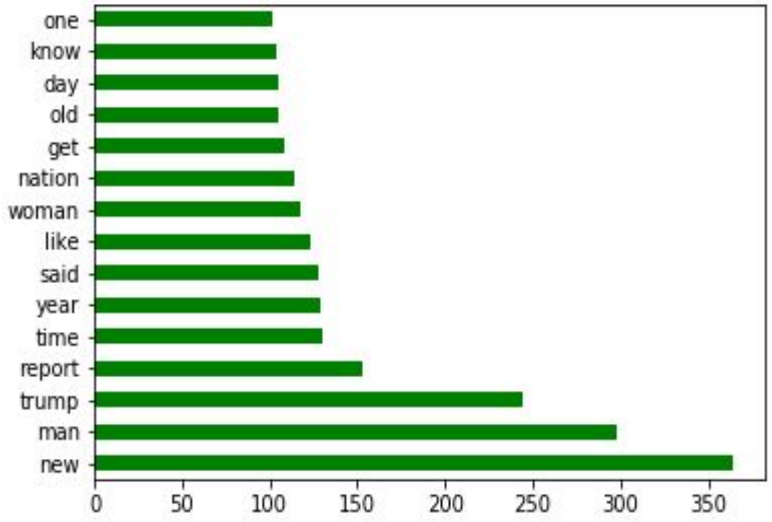
Opposing sides ---- Politics aside, often too sparse

What I picked:

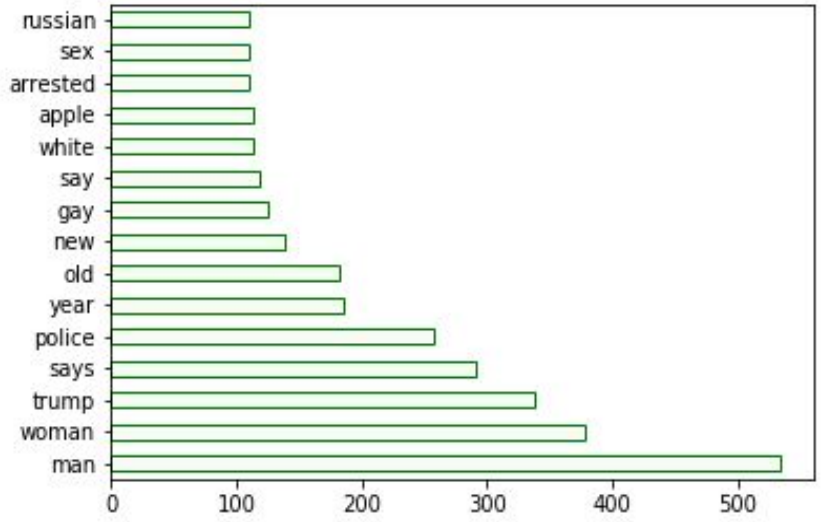
> *America's Finest News Source.*

Contrast

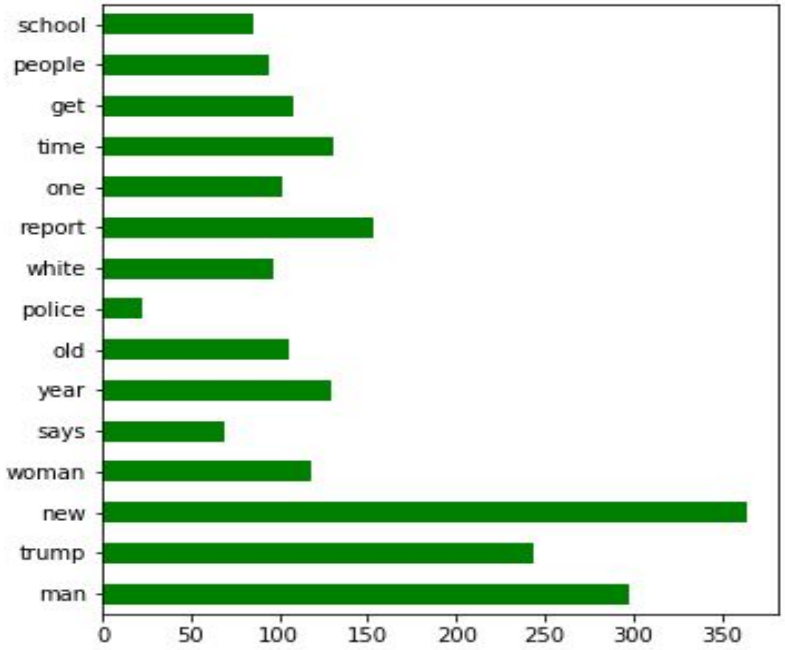
the ONION®

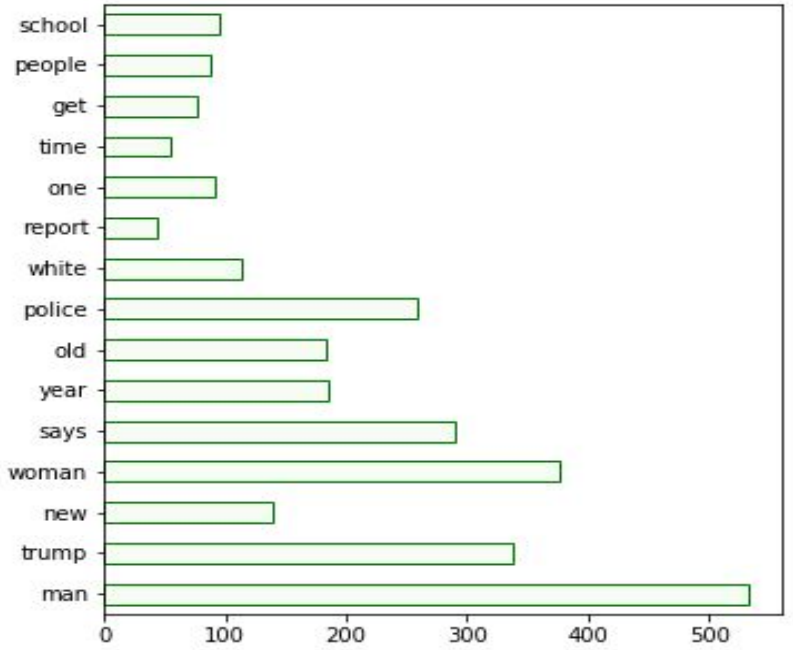


Not the ONION
REDDIT'S FINEST NEWS SOURCE

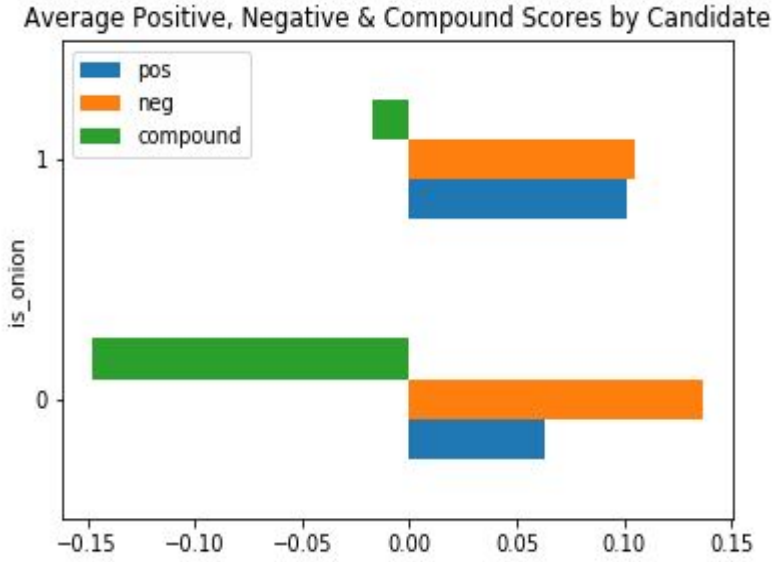


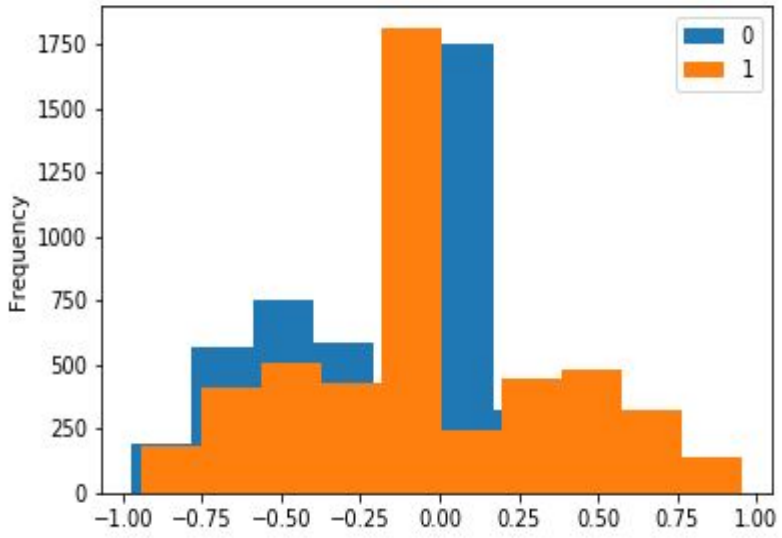
Compare

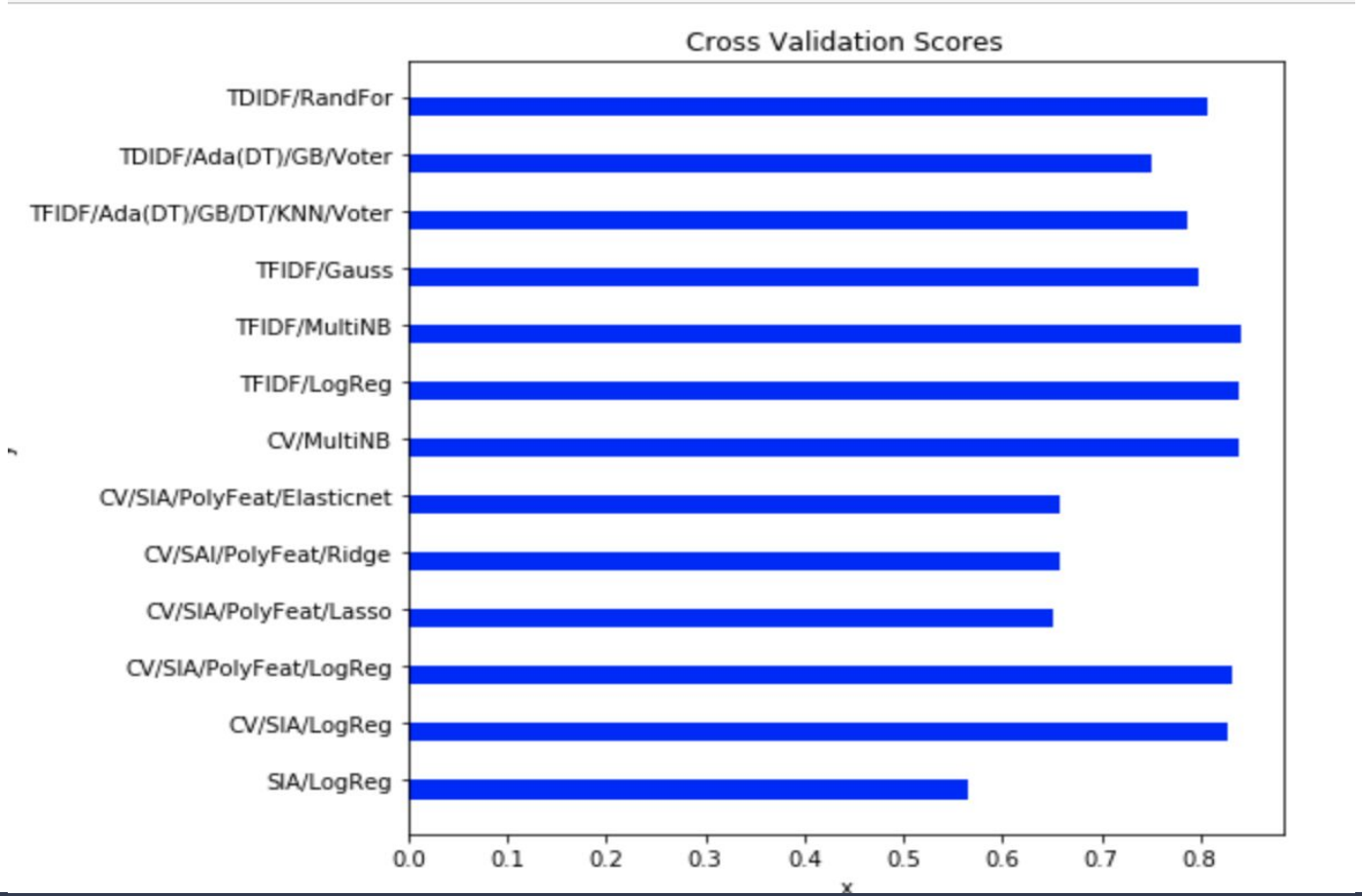

school
people
get
time
one
report
white
police
old
year
says
woman
new
trump
man
0
50
100
150
200
250
300
350


school
people
get
time
one
report
white
police
old
year
says
woman
new
trump
man
0
100
200
300
400
500

Vader Sentiment Intensity Analysis







Overall Results

By digits

method	cr_val_score
SIA/LogReg	0.5656
CV/SIA/LogReg	0.8283
CV/SIA/PolyFeat/LogReg	0.8333
CV/SIA/PolyFeat/Lasso	0.6520
CV/SAI/PolyFeat/Ridge	0.6572
CV/SIA/PolyFeat/Elasticnet	0.6572
CV/MultiNB	0.8386
TFIDF/LogReg	0.8386
TFIDF/MultiNB	0.8412
TFIDF/Gauss	0.7994
TFIDF/Ada(DT)/GB/DT/KNN/Voter	0.7872
TDIDF/Ada(DT)/GB/Voter	0.7516
TDIDF/RandFor	0.8072

Helpful if Poly Featuring a large dataset

```
model_cvec_feature_picker = ExtraTreesClassifier()
model_cvec_feature_picker.fit(df_X_train_cvec, y_train)
model_cvec_feature_picker.feature_importances_
feat_importances = pd.DataFrame(model_cvec_feature_picker.feature_importances_, index=
df_X_train_cvec.columns)
feat_importances.nlargest(25, columns= pd.RangeIndex(start=0, stop=1, step=1)     )
```

- Use instead of corr

Conclusion

Very good but short of outstanding results were obtained by using a model utilizing Multinomial Naive Bayes with TFIDF. Other models, including many that were more complex, were attempted, but meaningful gains were not made.