MISREAD

Fake News Classifier

Gene Woodstock

Data Scientist at

Reddit Integrity Team

Department Directors

Cross Discipline

Misinformation

Use machine learning algorithms to identify unreliable news postings and alert users of misleading or harmful misinformation being presented as fact.

How?

1. Gathed posts & comments from **r/TheOnion** and **r/worldnews** subreddits
2. Used classification algorithms to perform Natural Language Processing (NLP)
3. Calculated the likelihood a Reddit post contains misinformation
4. Categorized validity thresholds for news posts

* assumption: news posts in r/worldnews are factual

Data

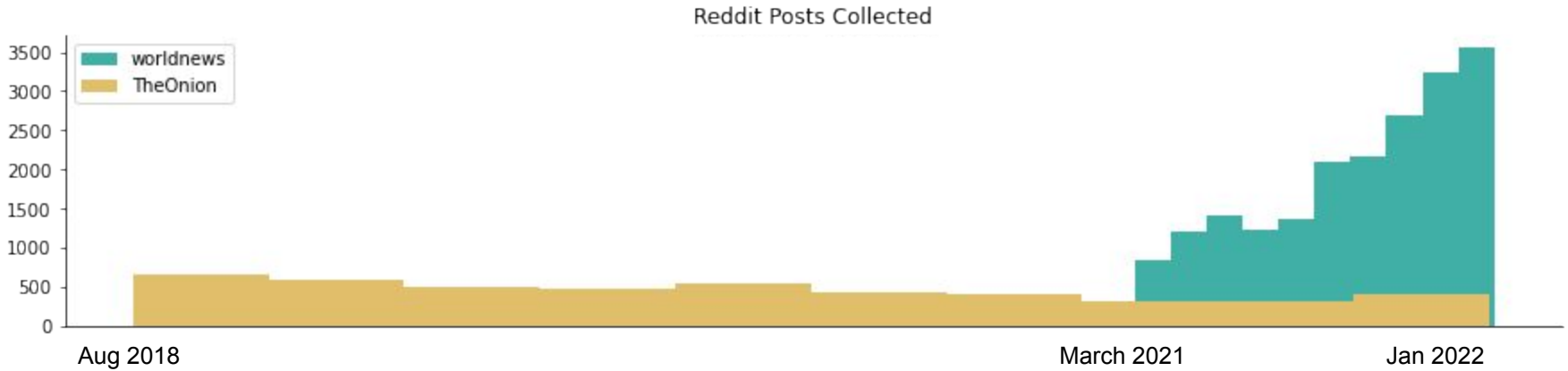
Collected via PushShift API

r/TheOnion

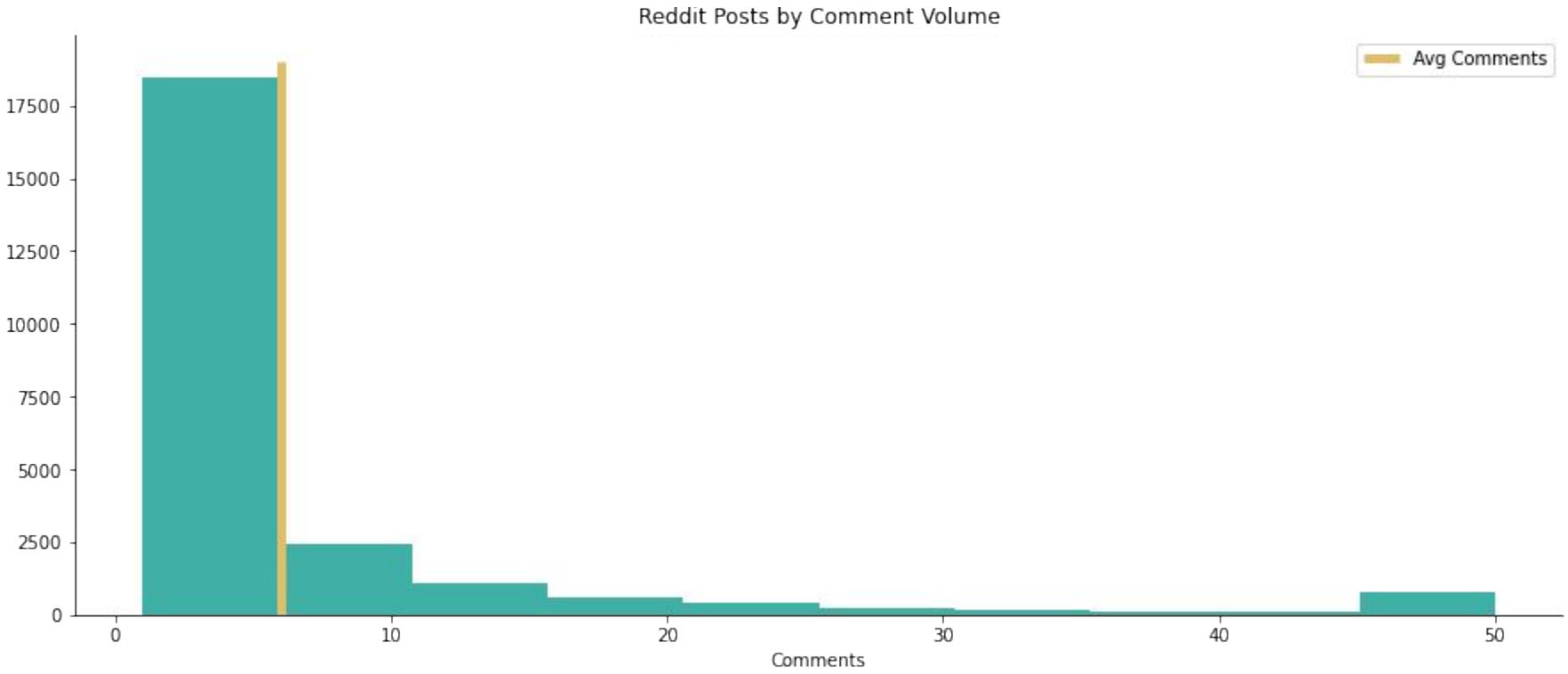
- 5,000 posts
- 35,000 comments
- Aug 2018 ~ Jan 2022
- Max 50 comments per post

r/worldnews

- 20,000 posts
- 180,000 comments
- March 2021 ~ Jan 2022
- Max 50 comments per post



Comments


Reddit Posts by Comment Volume
Avg Comments
17500
15000
12500
10000
7500
5000
2500
0
0
10
20
30
40
50
Comments

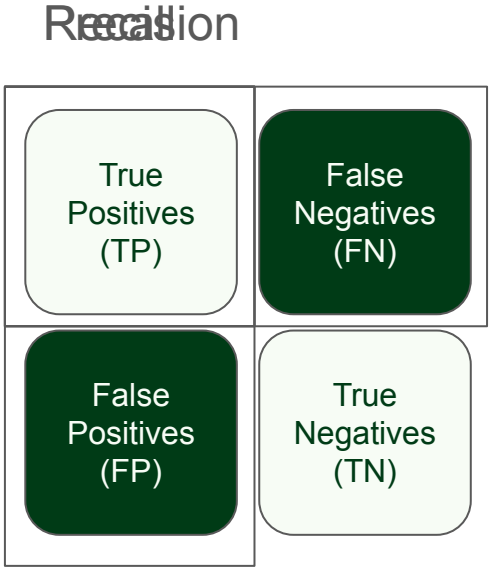
Success Metrics

Target:

- r/TheOnion

Measures:

- Optimize True Positives
- Reduce False Negatives
- Reduce False Positives



Flow 1

Count Vectorizer

9,563 words

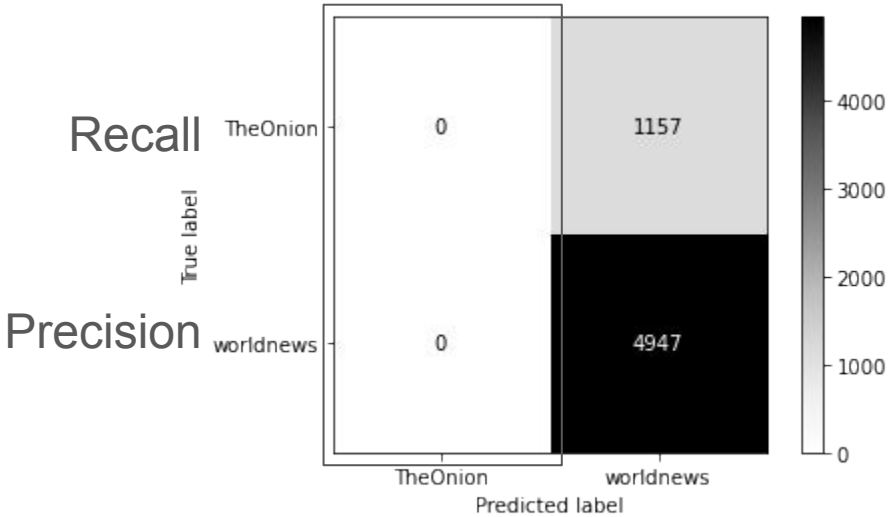
6 comments per post average

Model Selection

Compare:

- K Nearest Neighbors
- Logistic Regression
- Naive Bayes
- Decision Trees
- Random Forest
- Extra Trees

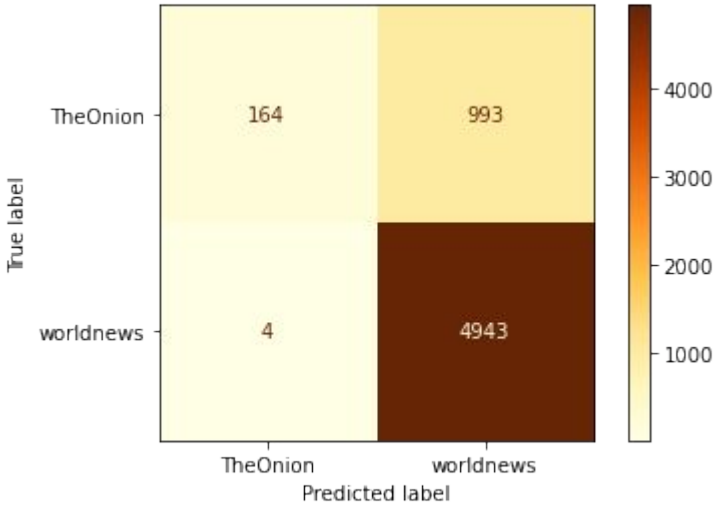
Baseline Acc: 81%



*Ada Boost, Gradient Boost, and XG Boost were excluded as they are too computationally expensive

Unsuccessful

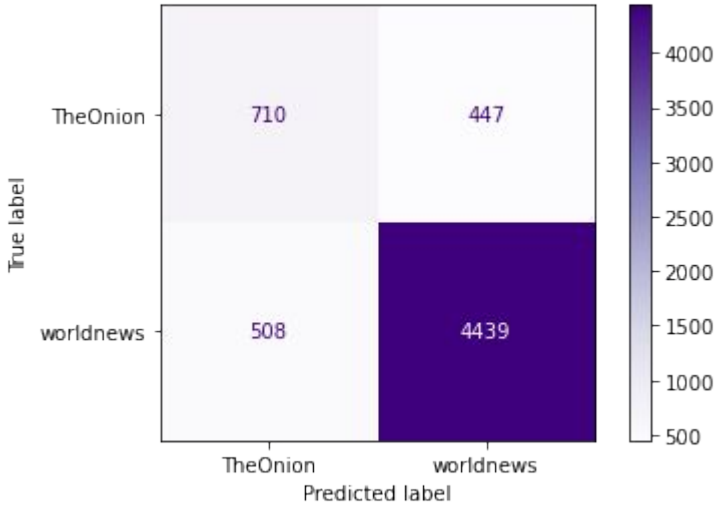
KNN



Recall: 14%

Precision: 98%

Decision Trees

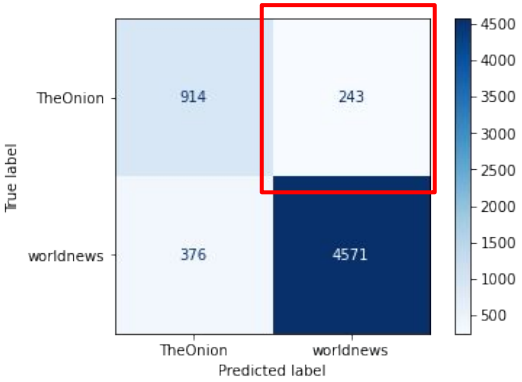


Recall: 61%

Precision: 58%

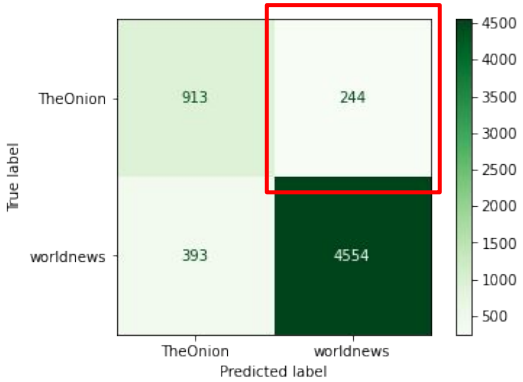
Successful

Logistic Regression



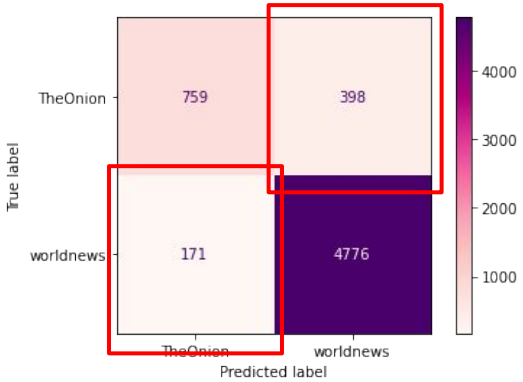
Recall: 79%
Precision: 71%

Naive Bayes

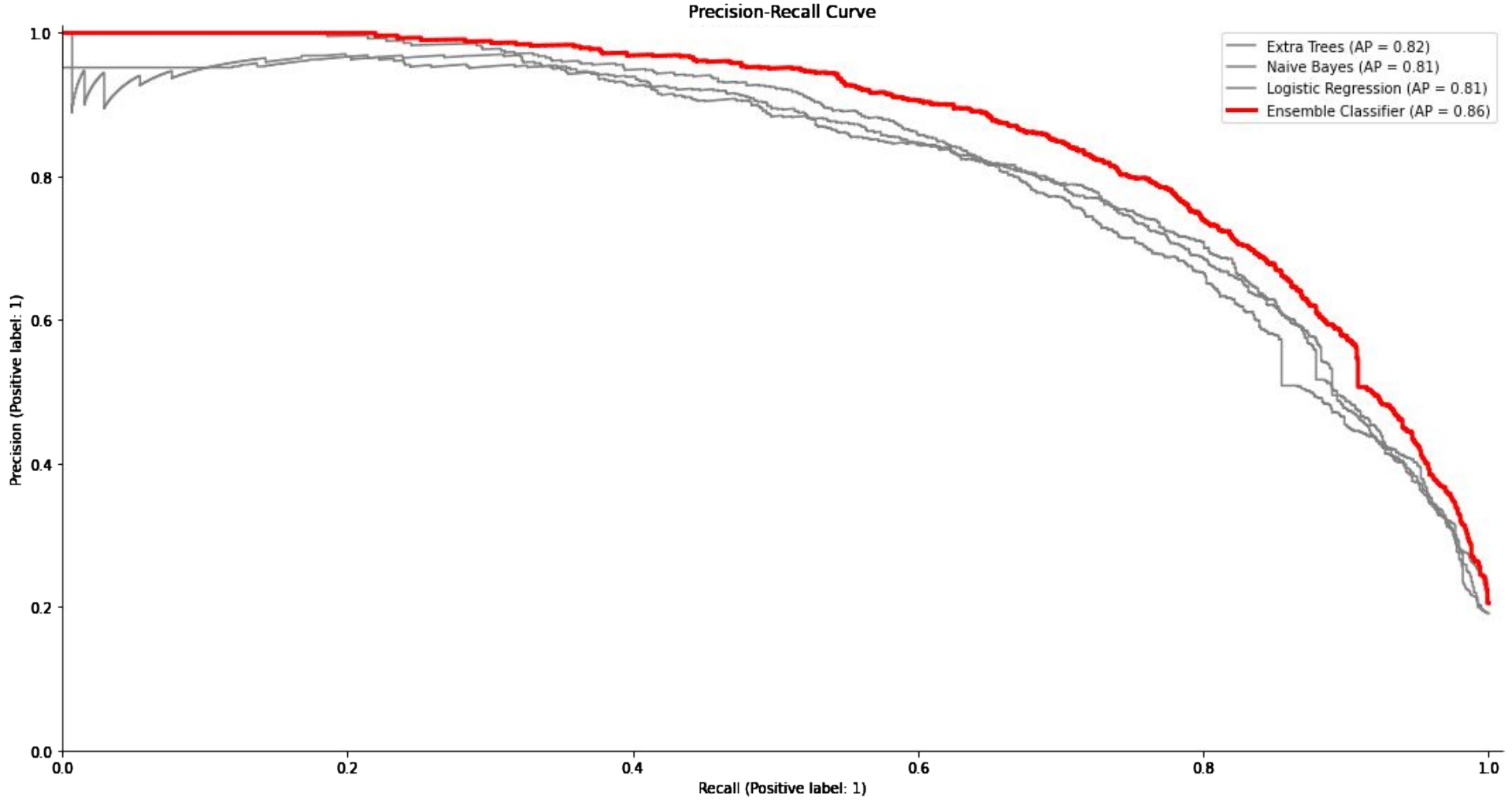


Recall: 79%
Precision: 70%

Extra Trees



Recall: 66%
Precision: 82%

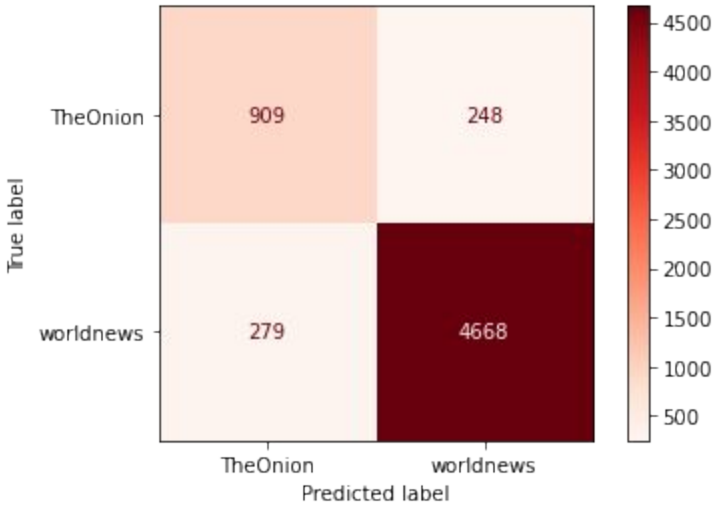

Precision-Recall Curve
Extra Trees (AP = 0.82)
Naive Bayes (AP = 0.81)
Logistic Regression (AP = 0.81)
Ensemble Classifier (AP = 0.86)
Precision (Positive label: 1)
Recall (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0

Ensemble

⅓ Logistic

⅓ Bayes

⅓ X-Trees



Recall: 80%
Precision: 77%
Accuracy: 91%
(+10% baseline)

Flow 2

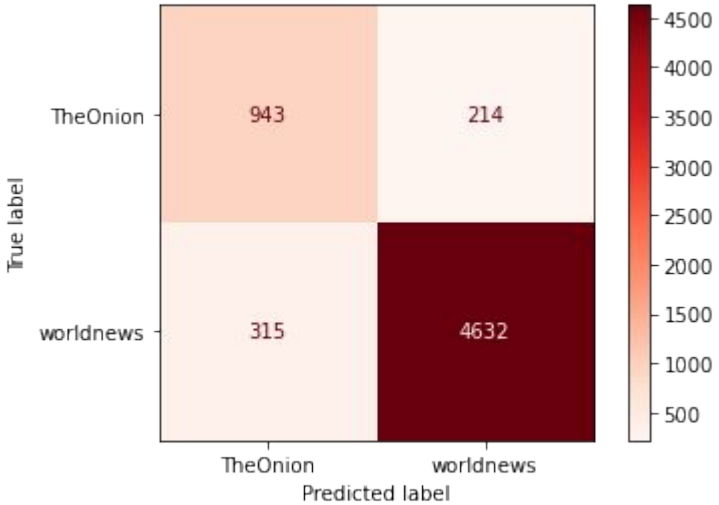
Term Frequency &
Inverse Document Frequency
vs
Count Vectorizer

10,958 words

6 comments per post average

CV vs TF-IDF

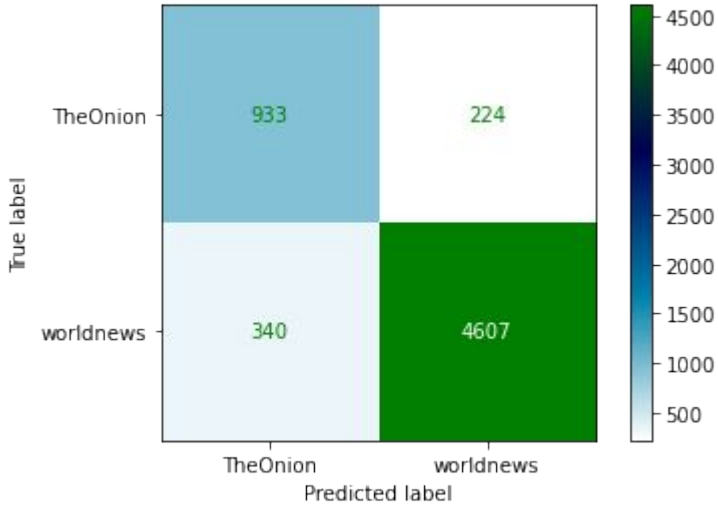
Count Vectorizer



Recall: 82%

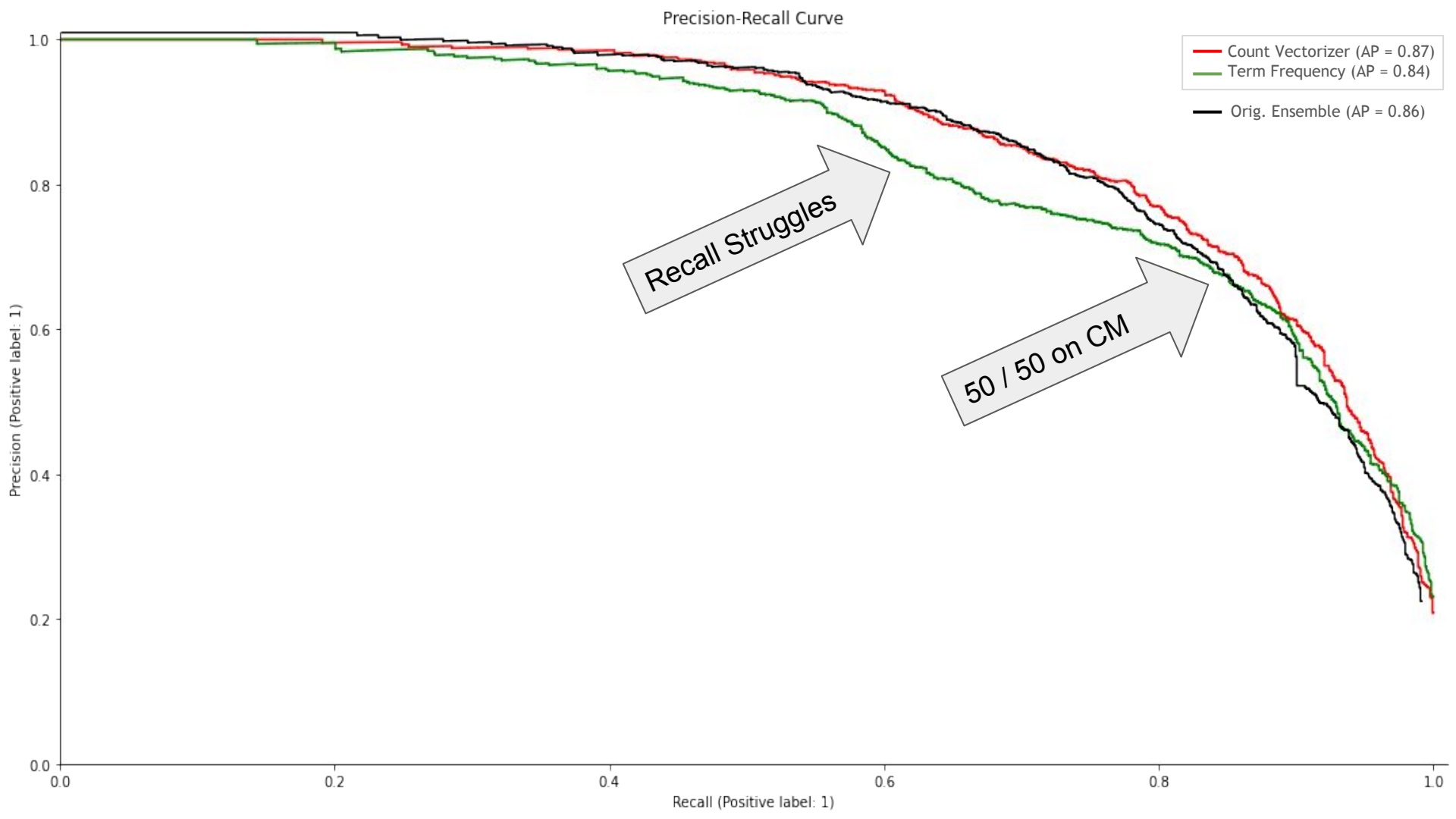
Precision: 75%

Term Frequency & Inverse Doc Frequency



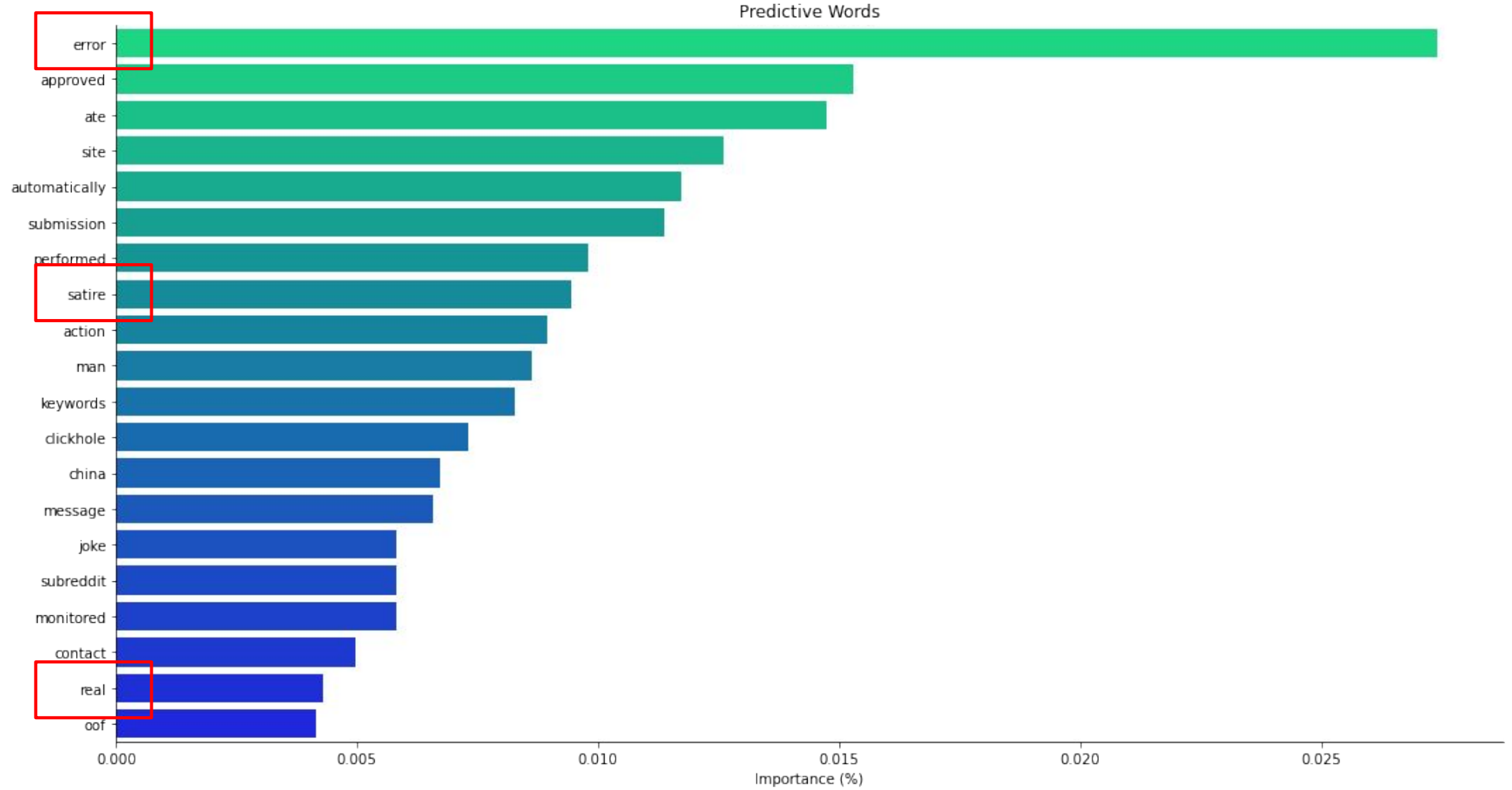
Recall: 81%

Precision: 73%


Precision-Recall Curve
Count Vectorizer (AP = 0.87)
Term Frequency (AP = 0.84)
Orig. Ensemble (AP = 0.86)
Recall Struggles
50 / 50 on CM
Precision (Positive label: 1)
Recall (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0

Words

Wisdom of the crowd


Predictive Words
error
approved
ate
site
automatically
submission
performed
satire
action
man
keywords
clickhole
china
message
joke
subreddit
monitored
contact
real
oof
0.000
0.005
0.010
0.015
0.020
0.025
Importance (%)

Opportunities

Generate misinformation predictions

Build “fake news” filter

Apply different UX treatments based on predictions thresholds

Misleading Post

Misleading Post

Threshold 1

75-85%

Misleading Post

Threshold 2
85% and up

Model Prediction: 87.7% Fake News

Takeaways

Recall Rate - 80%
Precision Rate - 75%

Ensemble classification identifies posts likely to contain misinformation

Fake news filter informs users of harmful content

THANK YOU

Questions?

What’s Next?

- Undersample majority class
- Bootstrap minority class
- Bi-gram and tri-gram features
- Boosting model techniques
- Sentiment analysis
- Factor in meta-data to model
- More data!!