Predicting Comment Volume for News Articles

Metis

Natural Language Processing and Unsupervised Learning

Jonathon Bowyer

Introduction and Domain

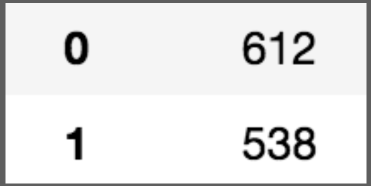
- NYT News Articles

- Classify articles as high-comment or low-comment
  - Better engagement
  - Increase repeat customers

- Determine high-comment topics
  - Placement of articles
  - Number of articles per topic

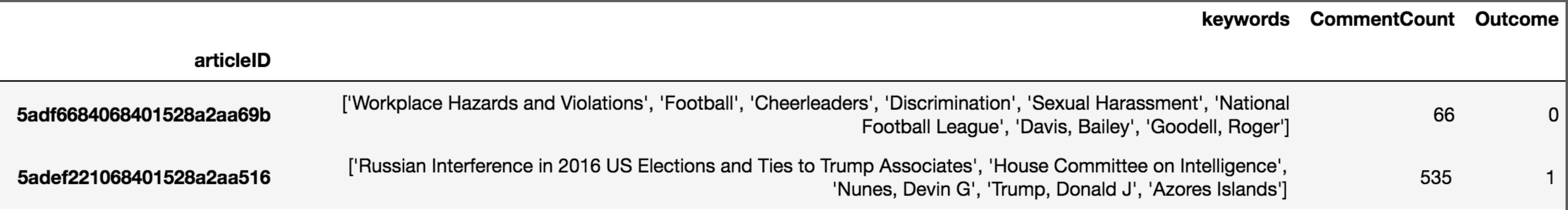
Data

- 2 Kaggle datasets: 1) NYT Articles & 2) Comments on articles
  - ~1,200 Articles from April 2018
- Utilized keywords from datasets
- Target: article received above 100 comments
  - (0 = below, 1 = above)
- Fairly balanced dataset:



0	612
1	538

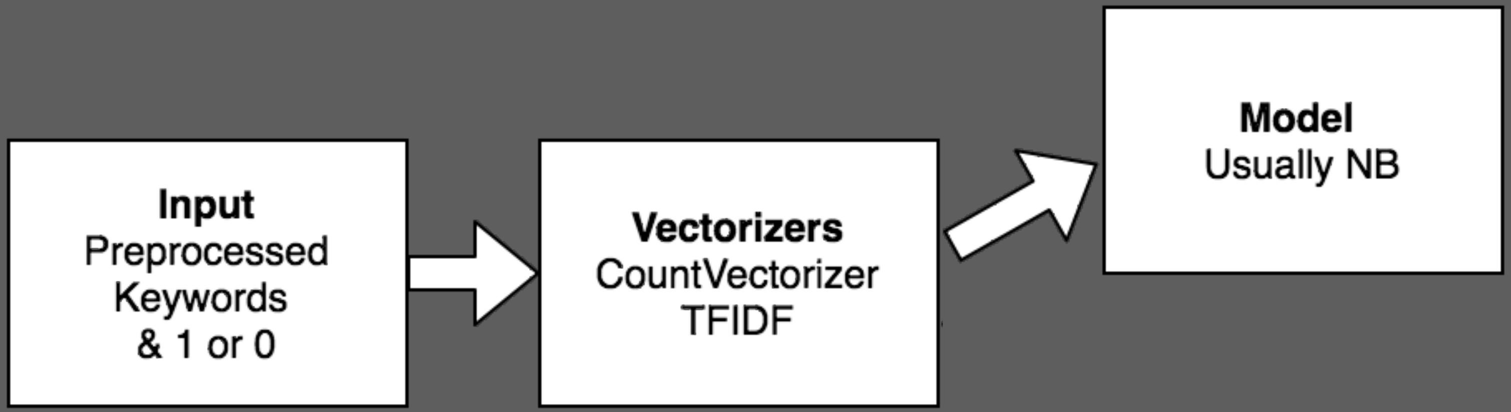
Data Cont.



	keywords	CommentCount	Outcome
articleID			
5adf6684068401528a2aa69b	['Workplace Hazards and Violations', 'Football', 'Cheerleaders', 'Discrimination', 'Sexual Harassment', 'National Football League', 'Davis, Bailey', 'Goodell, Roger']	66	0
5adef221068401528a2aa516	['Russian Interference in 2016 US Elections and Ties to Trump Associates', 'House Committee on Intelligence', 'Nunes, Devin G', 'Trump, Donald J', 'Azores Islands']	535	1

NLP Modeling

NLP Modeling



Input
Preprocessed
Keywords
& 1 or 0

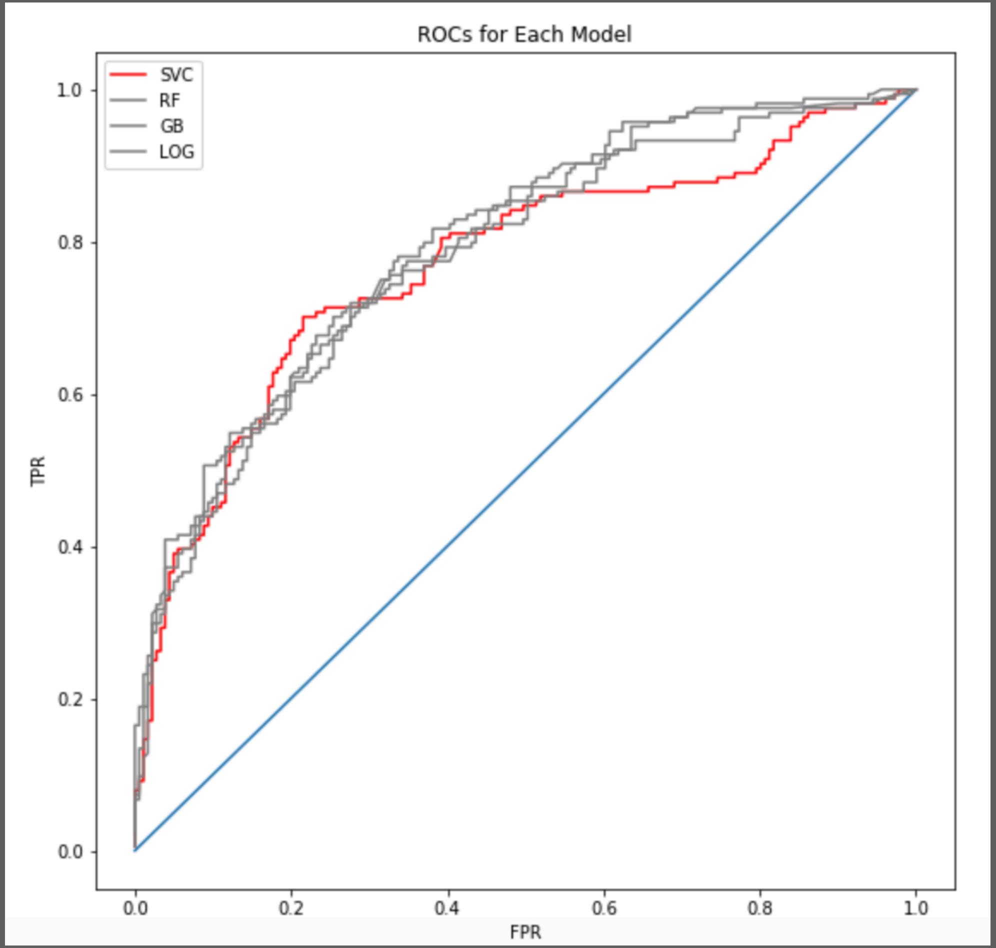
Vectorizers
CountVectorizer
TFIDF

Model
Usually NB

Curse of dimensionality: overfitting

Train: 89% accuracy
Test: 69% accuracy

Modeling

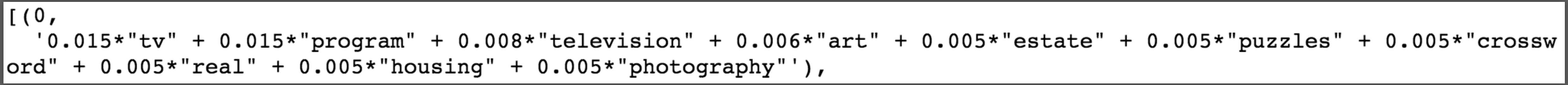


Models similar and successful: train/test ~77%

Recommendation:
SVM, minimizing FPR

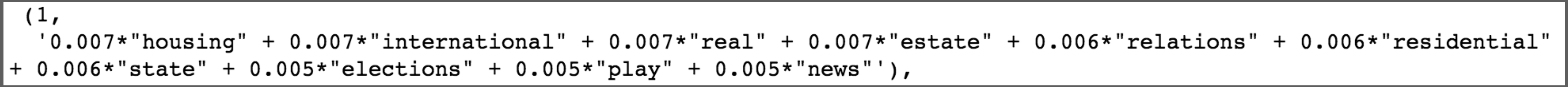
Topics in Low Comment Articles

Cultural (tv, art, photography, crosswords)



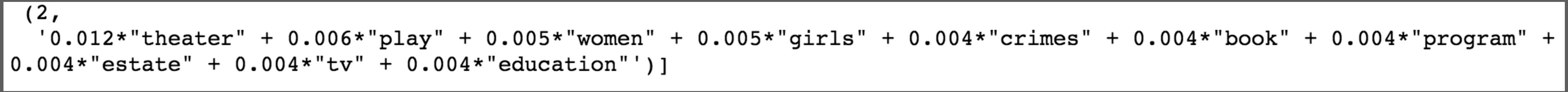
```
[(0,
  '0.015*"tv" + 0.015*"program" + 0.008*"television" + 0.006*"art" + 0.005*"estate" + 0.005*"puzzles" + 0.005*"crossw
ord" + 0.005*"real" + 0.005*"housing" + 0.005*"photography"'),
```

Real estate



```
 (1,
  '0.007*"housing" + 0.007*"international" + 0.007*"real" + 0.007*"estate" + 0.006*"relations" + 0.006*"residential"
+ 0.006*"state" + 0.005*"elections" + 0.005*"play" + 0.005*"news"'),
```

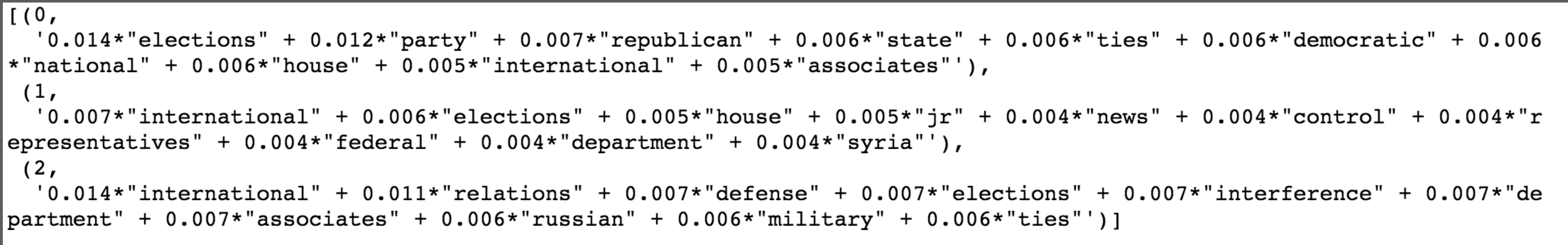
Women + culture



```
 (2,
  '0.012*"theater" + 0.006*"play" + 0.005*"women" + 0.005*"girls" + 0.004*"crimes" + 0.004*"book" + 0.004*"program" +
0.004*"estate" + 0.004*"tv" + 0.004*"education"')]
```

Topics in High Comment Articles

Politics: Trump (stop word), elections, international relations, military



```
[(0,
  '0.014*"elections" + 0.012*"party" + 0.007*"republican" + 0.006*"state" + 0.006*"ties" + 0.006*"democratic" + 0.006
*"national" + 0.006*"house" + 0.005*"international" + 0.005*"associates"'),
 (1,
  '0.007*"international" + 0.006*"elections" + 0.005*"house" + 0.005*"jr" + 0.004*"news" + 0.004*"control" + 0.004*"r
epresentatives" + 0.004*"federal" + 0.004*"department" + 0.004*"syria"'),
 (2,
  '0.014*"international" + 0.011*"relations" + 0.007*"defense" + 0.007*"elections" + 0.007*"interference" + 0.007*"de
partment" + 0.007*"associates" + 0.006*"russian" + 0.006*"military" + 0.006*"ties"')]
```

Conclusion + Future Work

Conclusion

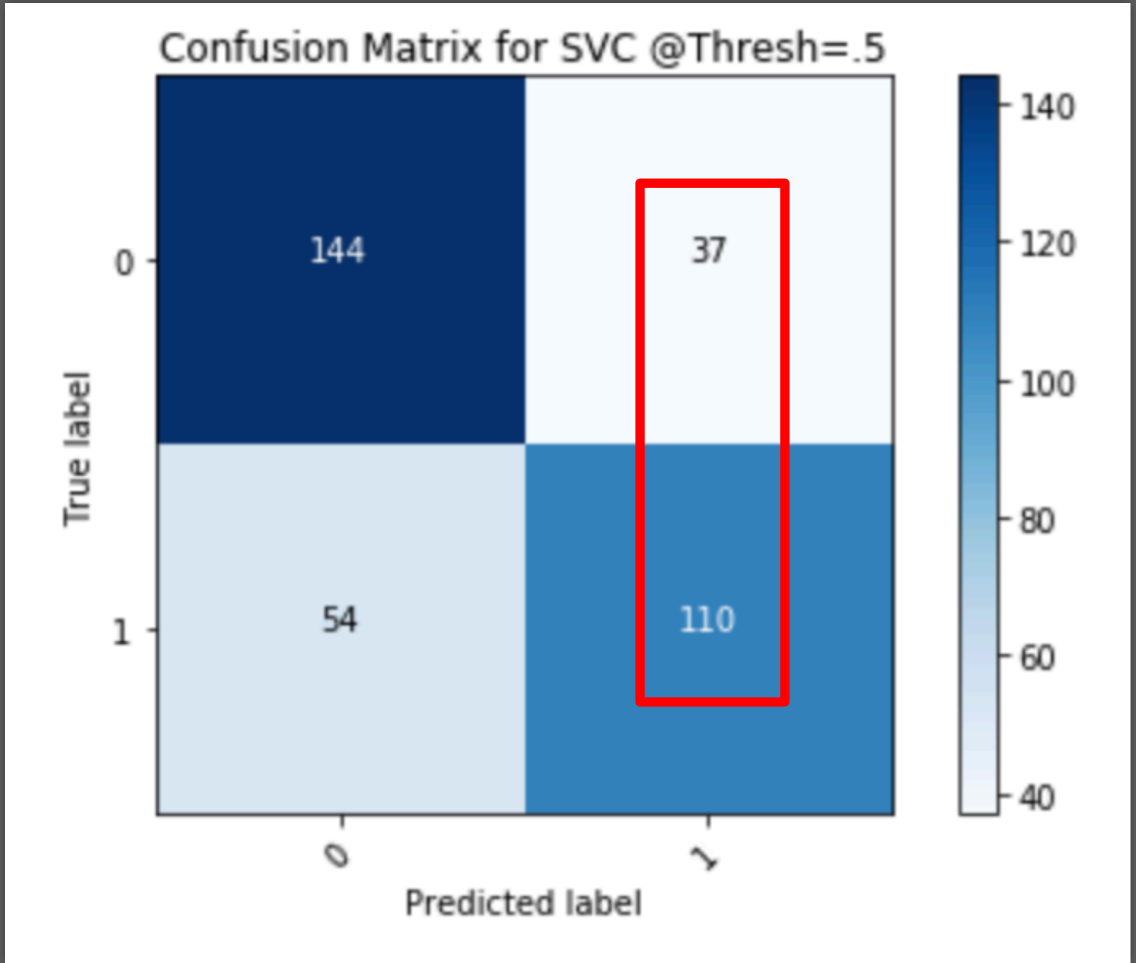
- Can accurately predict high-comment articles
- Have topics readers are interested in

Future work

- Add more data (more months)
  - Weight older articles as less important
- K-means clustering for these topics (similar to topic modeling)

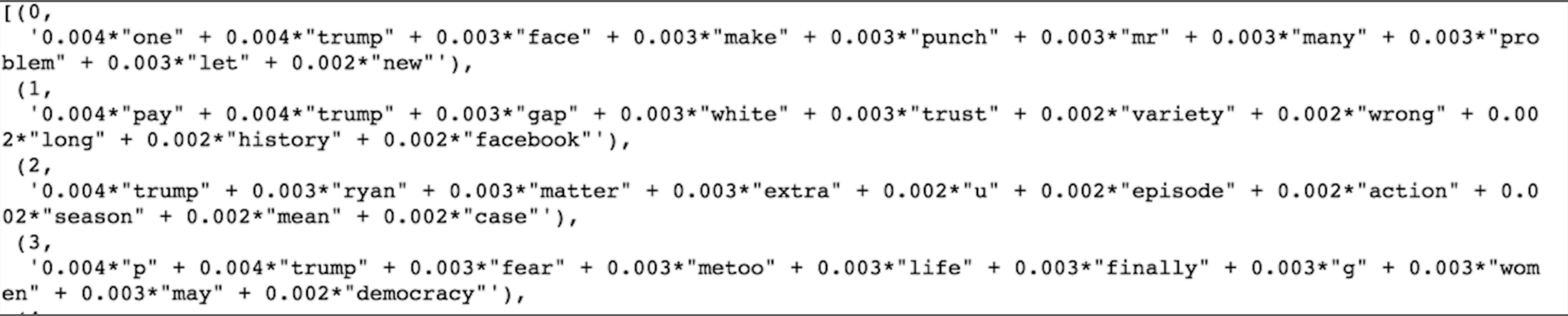
Appendix

SVM Confusion Matrix



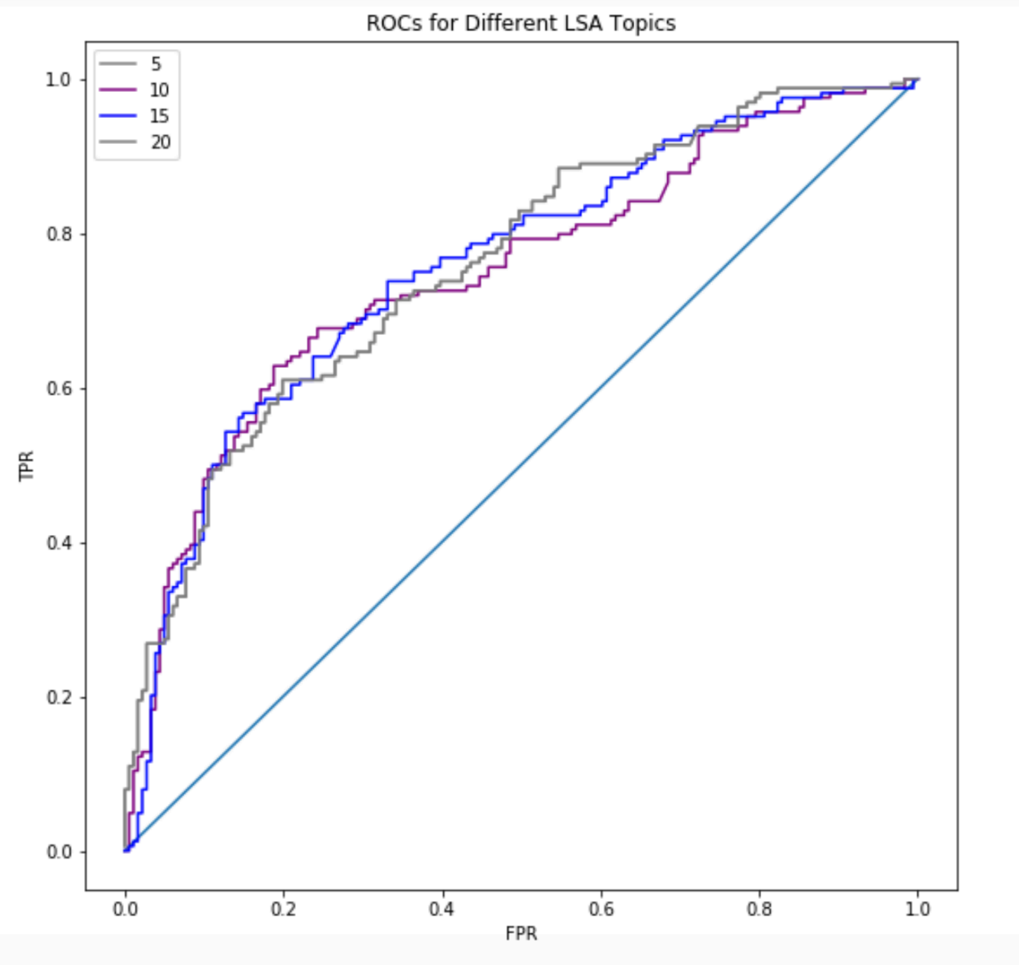
Headlines

- Headlines were way less predictive, in fact most models went below 50%
- Topics were odd
- Headlines aim to be unique and catchy, less able to sort through actual meaning/connection to what the article is about. Worse proxy

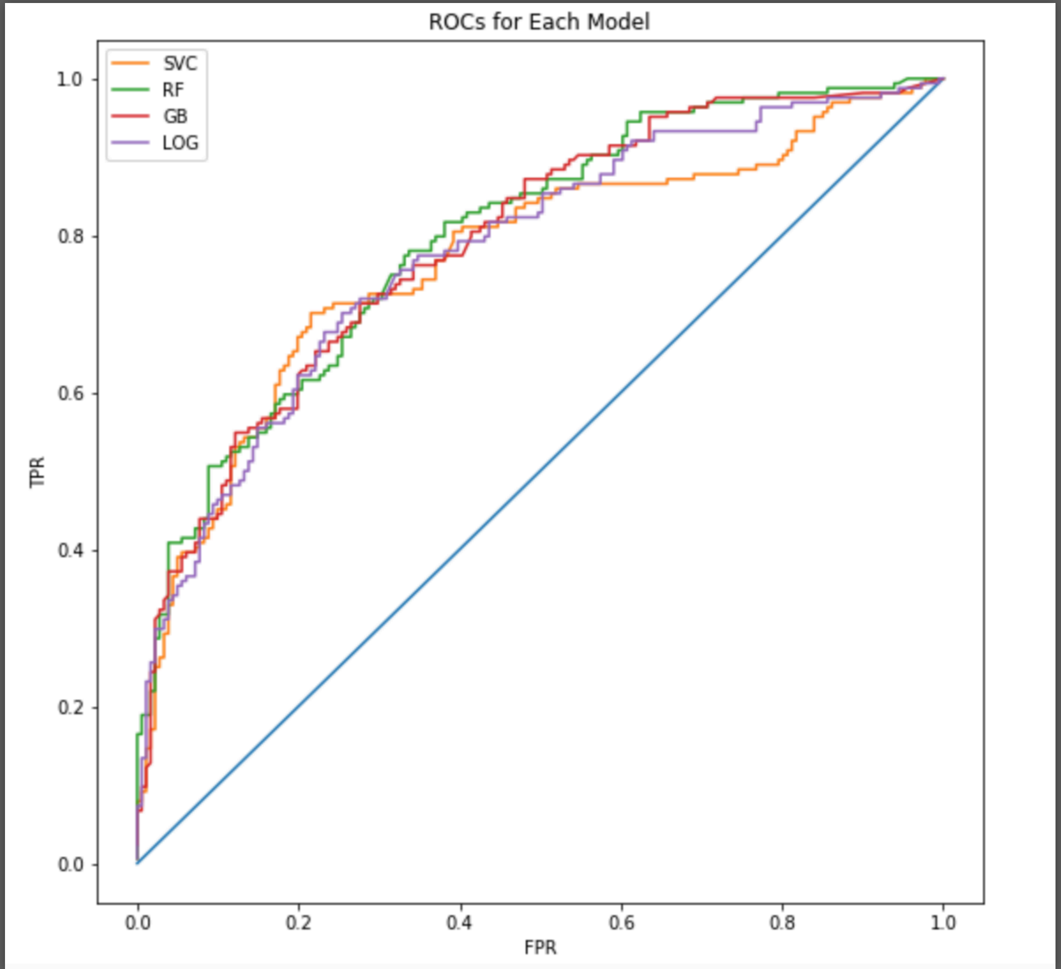


```
[(0,
  '0.004*"one" + 0.004*"trump" + 0.003*"face" + 0.003*"make" + 0.003*"punch" + 0.003*"mr" + 0.003*"many" + 0.003*"pro
blem" + 0.003*"let" + 0.002*"new"'),
 (1,
  '0.004*"pay" + 0.004*"trump" + 0.003*"gap" + 0.003*"white" + 0.003*"trust" + 0.002*"variety" + 0.002*"wrong" + 0.00
2*"long" + 0.002*"history" + 0.002*"facebook"'),
 (2,
  '0.004*"trump" + 0.003*"ryan" + 0.003*"matter" + 0.003*"extra" + 0.002*"u" + 0.002*"episode" + 0.002*"action" + 0.0
02*"season" + 0.002*"mean" + 0.002*"case"'),
 (3,
  '0.004*"p" + 0.004*"trump" + 0.003*"fear" + 0.003*"metoo" + 0.003*"life" + 0.003*"finally" + 0.003*"g" + 0.003*"wom
en" + 0.003*"may" + 0.002*"democracy"'),
```

LSA Topic Number



Modeling with colors



Sentiment analysis trouble

- Article: President Trump welcomed President Emmanuel Macron of France to the White House on Tuesday for the first state visit since taking office.

- Comments:
- Macron, and every other world leader, must be aware that whatever agreement they think they come to, can and will be undone by the toxic personalities on Fox who Trump gives more credibility to than our intelligence agencies. Mr. Macron will be embarrassed by Trump, sure as everyone else in his coterie has been. Sentiment (**polarity=0.28125**, subjectivity=0.5034722222222222)
- Enjoy the Enfant Terrible, Mr. Macron.<br/><br/>Don't forget to change his dirty diapers. Sentiment (**polarity=-0.39999999999999997**, subjectivity=0.7666666666666666)

- Sentiment will be negative if: against Trump, against Trump meeting with Macron, dislike the article (how it's written), etc. Too many possible attributions