Fake News Project

Motivation For Project

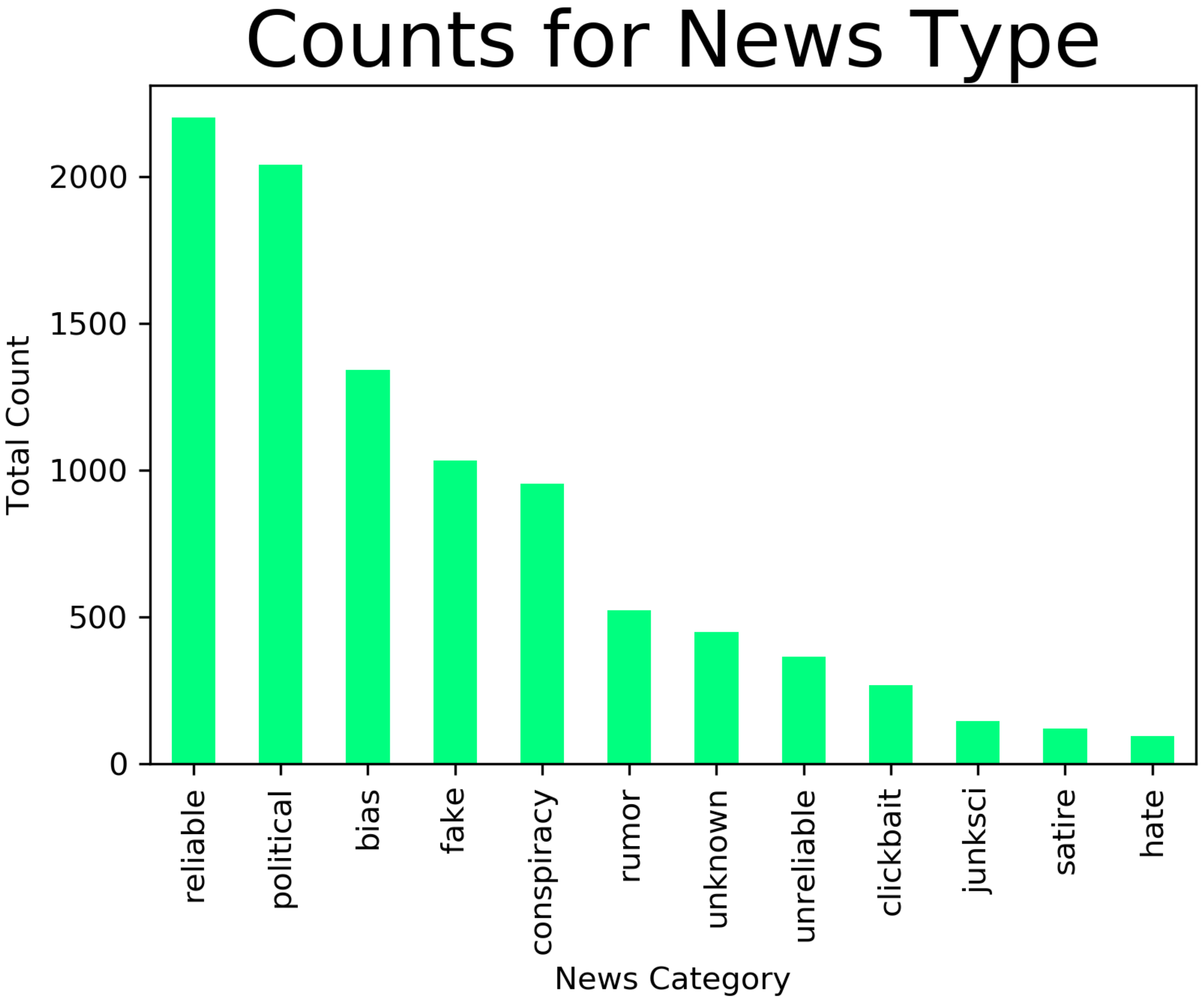
- Fake news detection is important for several industries.
- Notably, regulatory agencies have begun to target websites with user-generated content (such as social media sites) for allowing the circulation of fake news on their platforms.
- Automated fake news detection can be useful for major websites to identify potentially problematic content in advance.

Questions for Analysis

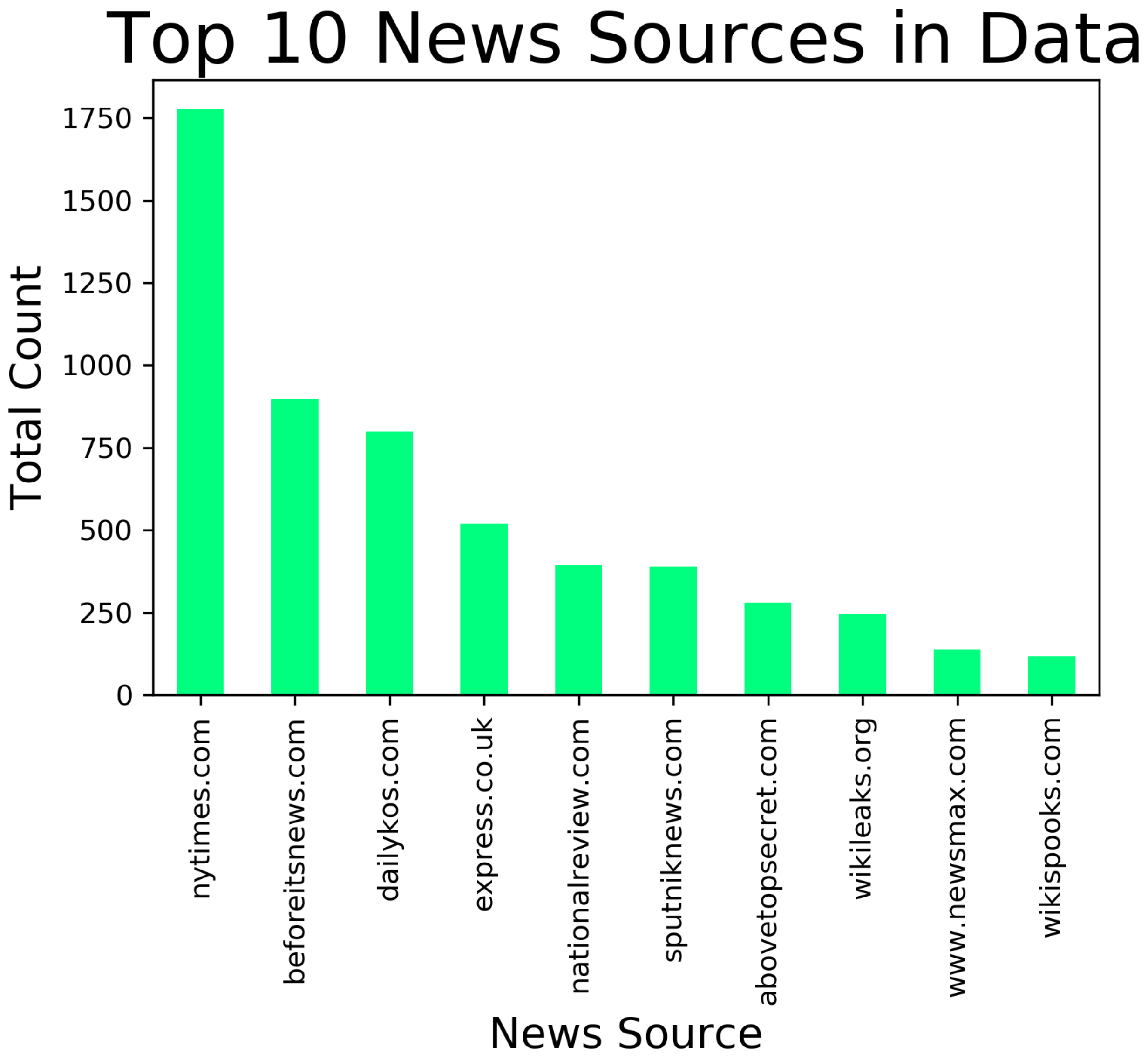
- What can machine learning techniques tell us about fake news detection?

- How do different styles of sampling data influence results?

Category Counts

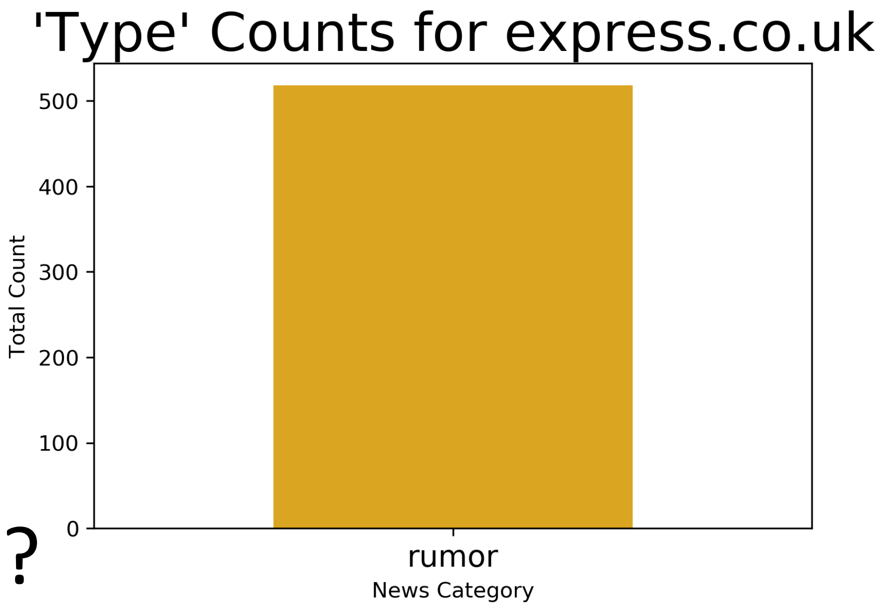


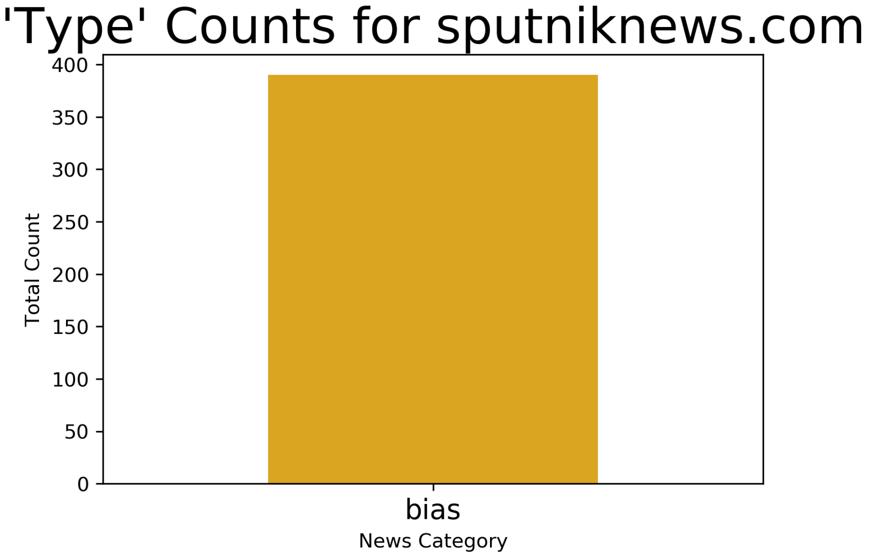
News Source Counts

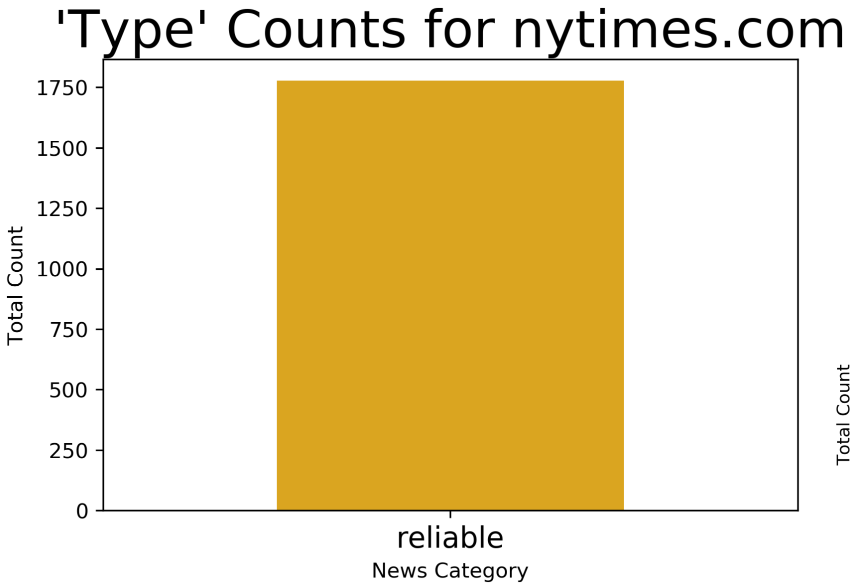


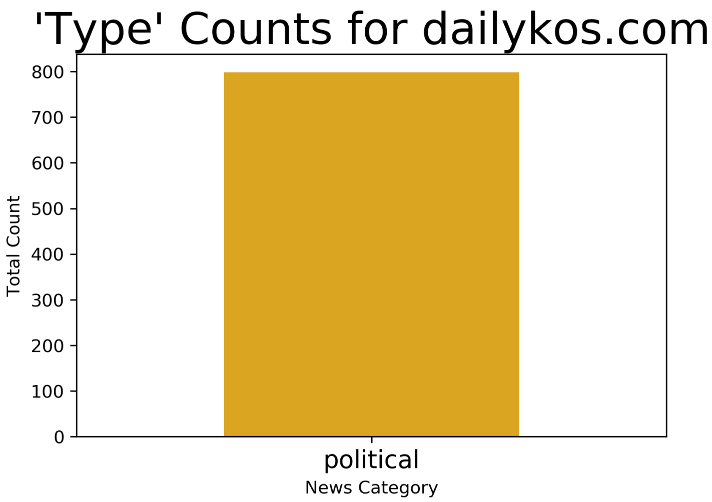
Every single article
from a given source
is in one category
in the data

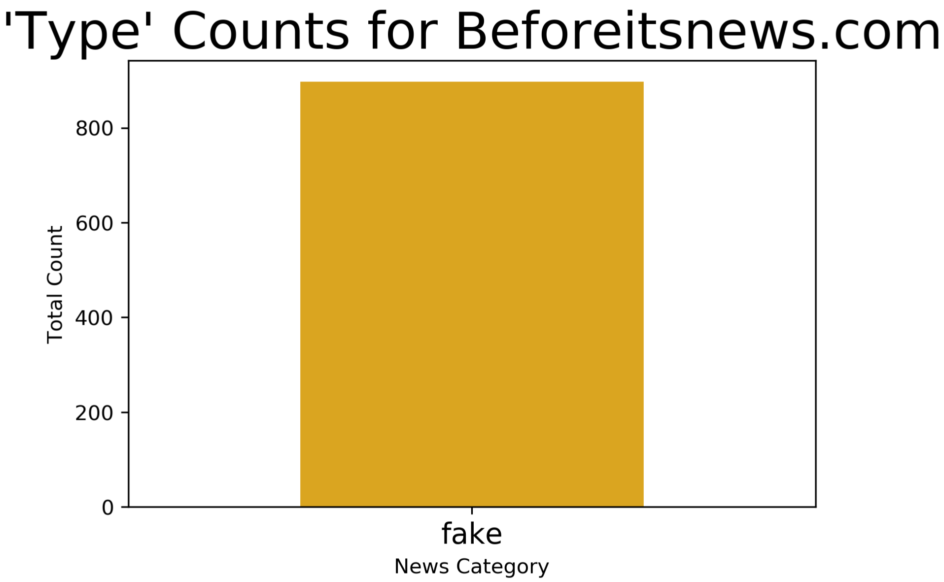
Is this problematic?



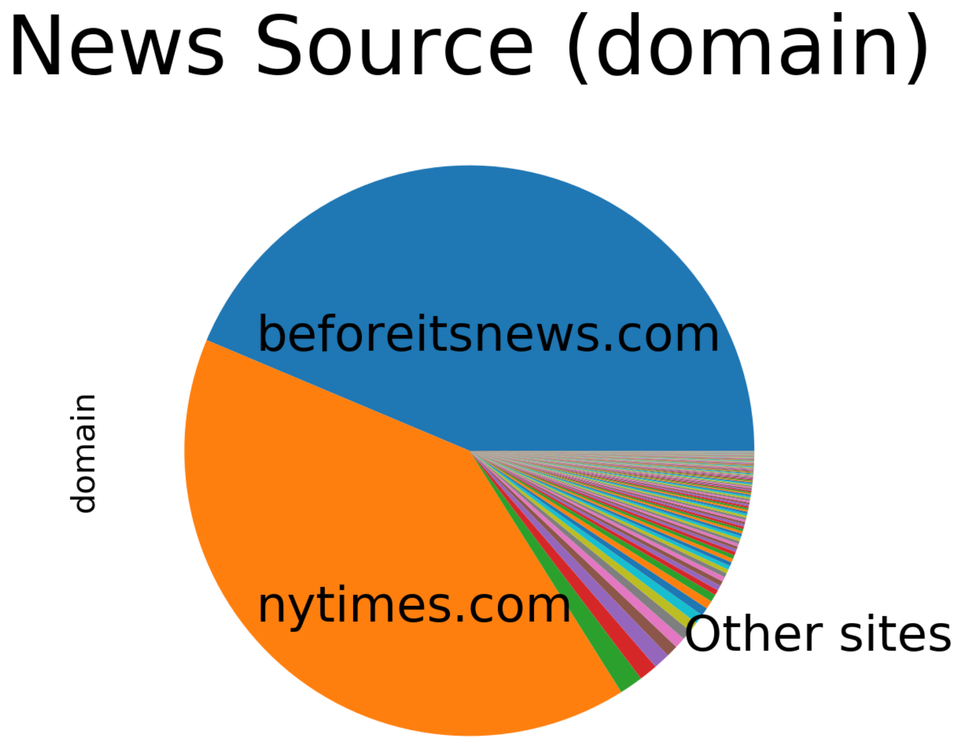


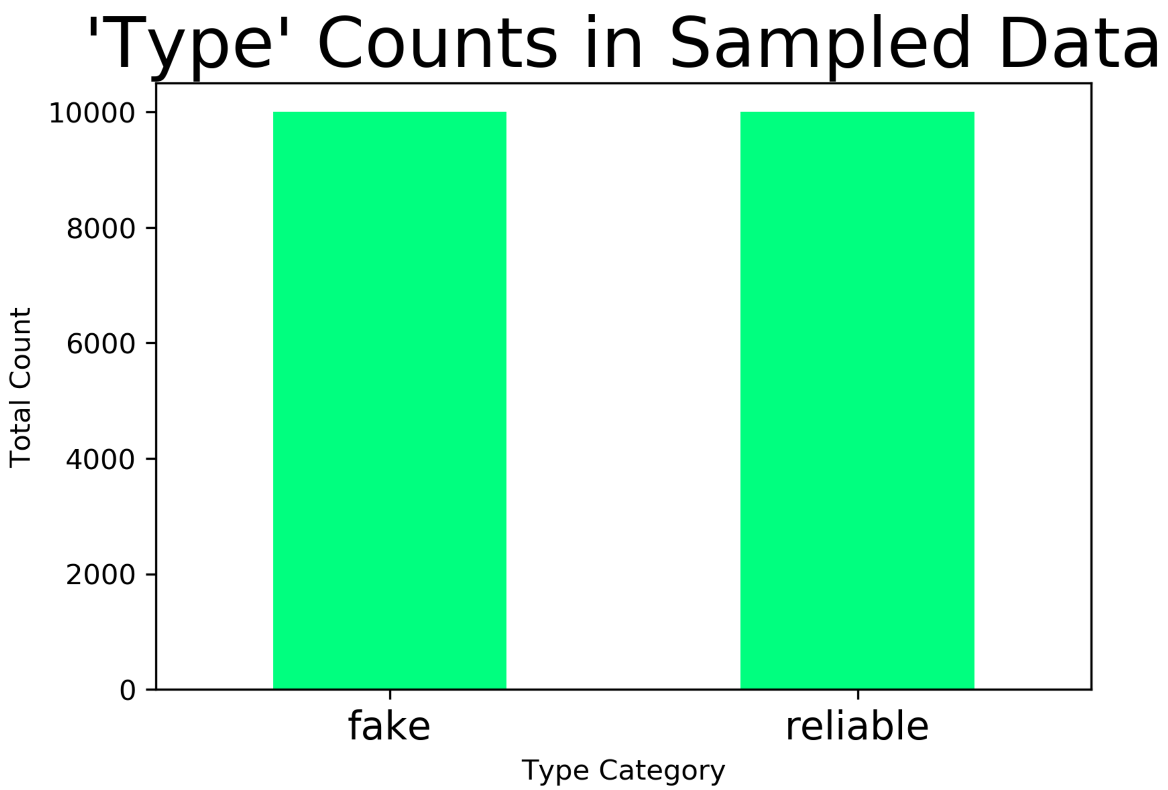






Initial Analysis – Reliable vs. Fake





Sentiment Analysis - Polarity

	SIA Polarity Score		**TextBlob Polarity Score**	
	Labelled "Fake"	Labelled "Reliable"	Labelled "Fake"	Labelled "Reliable"
Percentage of documents with a score greater than 0.2 (positive)	59.39%	67.21%	17.49%	11.75%
Percentage of documents with a score less than -0.2 (negative)	34.64%	25.81%	13.3%	7.30%
Percentage of documents with a score between -0.2 and 0.2 (neutral)	5.97%	6.98%	81.*18%*	87.52%

More documents labelled "fake" had more negative sentiment for both SIA and TextBlob.

Some differences between SIA and TextBlob regarding high polarity scores.

Sentiment Analysis - Subjectivity

	TextBlob Subjectivity	
	Labelled “Fake”	Labelled “Reliable”
Percentage of documents with a score greater than 0.55 (subjective)	11.28%	7.62%
Percentage of documents with a score less than 0.45 (objective)	57.84%	68.00%
Percentage of documents with a score between 0.45 and 0.55 (neutral)	30.88%	24.38%

Predictive Modeling – Reliable vs. Fake

		Vectorization Technique		
		Bag of Words	Tf-idf	Tf-idf with two bigrams
Classifier	MultinomialNB()	86.7%	87.2%	90.4%
	LinearSVC()	87.0%	90.8%	91.7%
	XGB Classifier()	87.5%	89.0%	83.3%

Most Predictive features for Initial Reliable vs. Fake Analysis

6.2862	main stori
6.2685	read main
5.9580	advertis continu
5.6588	continu read
4.7699	new york
4.2447	to re
2.3737	an articl
1.9920	next in
1.7252	said would

- These results indicate that the fact that a given article is from the New York Times is more predictive than anything else in the data. Because of this, the data was resampled (see following slide)

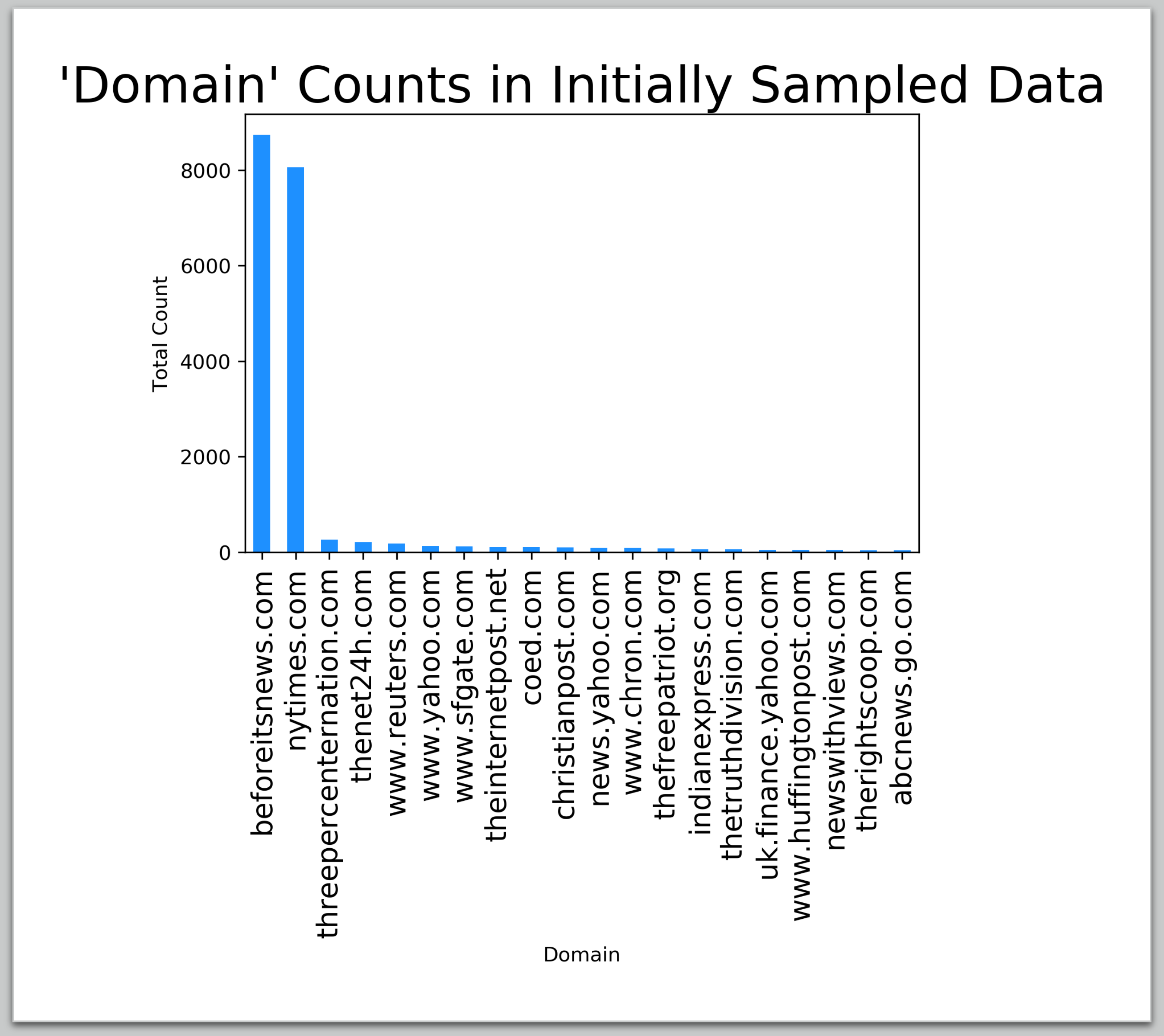
Topic Modeling

- Documents labelled “reliable” had two identifiable topics:
  - - Vocabulary linked to the New York Times
    - This indicates that all articles from given sources given one label is a significant issue in this dataset.
  - - Vocabulary linked to arts and performance
    - These may be articles from the arts section of The New York Times

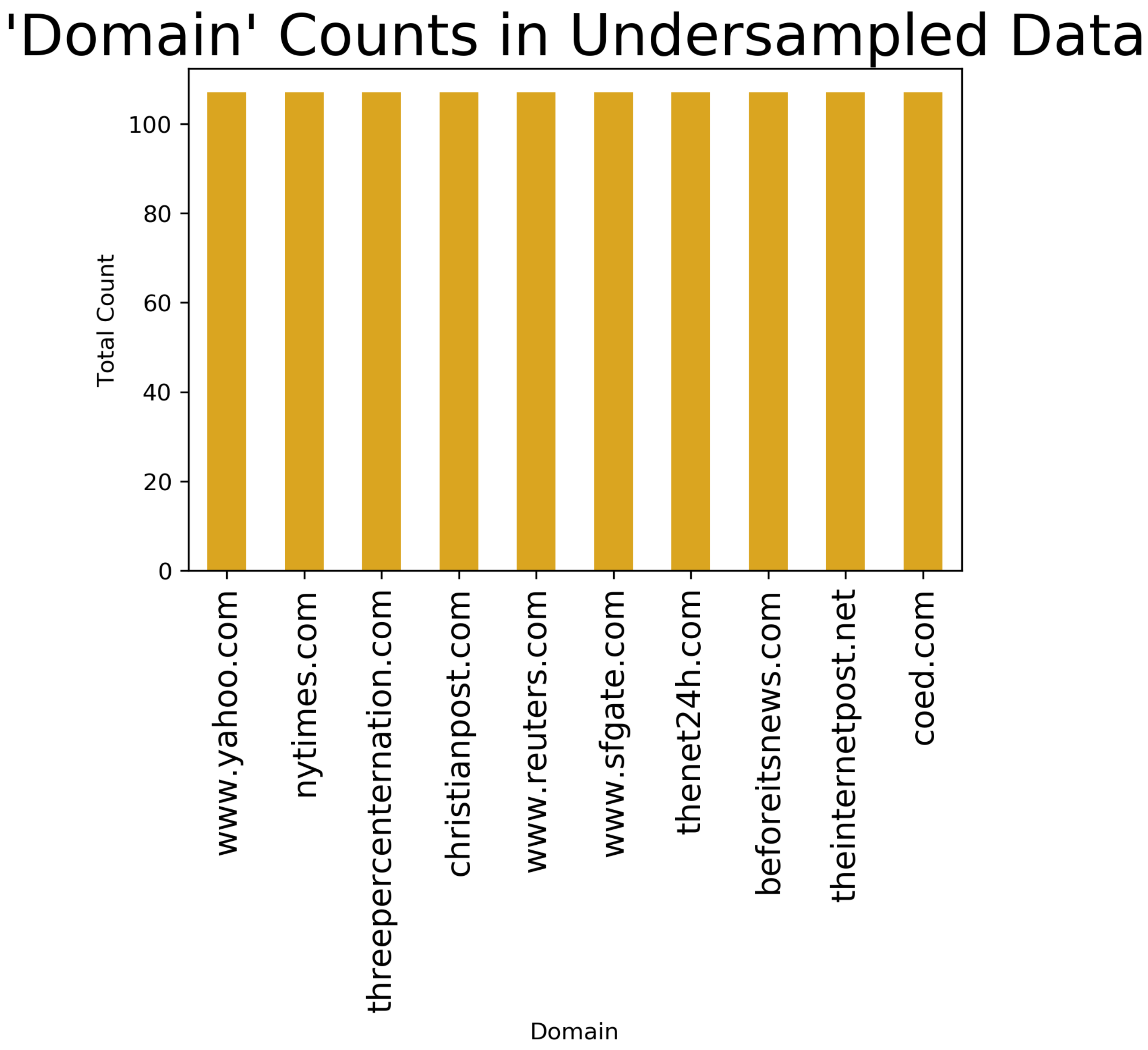
- Documents labelled “fake” had two identifiable topics:
  - - Religious vocabulary coinciding with the words “day” and “night”: (1, '-0.460*"christ" + -0.371*"day" + -0.220*"god" + -0.205*"jesu" + -0.172*"night”’)
  - Vocabulary linked to the ACA (Afforable Care Act) aka “Obamacare”: (2, '0.507*"obamacar" + 0.466*"obama" + -0.262*"market" + 0.238*"websit" + 0.186*"insur"')

Resampling

The New York Times and beforeitsnews.com were overrepresented in the initial sample



Data was resampled for better balance across domain



Lower accuracy than initial sampling. Classifiers from initial sub-sample were likely learning to predict majority classes

Predictive Accuracy – Resampled Data

		Vectorization Technique		
		Bag of Words	Tf-idf	Tf-idf with two bigrams
Classifier	MultinomialNB()	77.6%	72.3%	77.6%
	LinearSVC()	86.4%	89.0%	84.1%
	XGB Classifier()	84.7%	87.6%	77.6%

Predictive Features

Unigrams

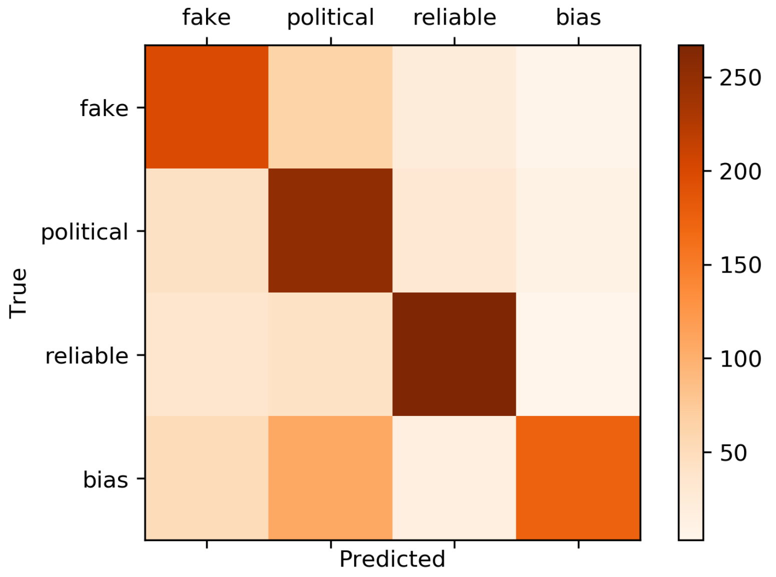
2.8374	2016
2.3298	ap
2.2535	nov
2.1699	november
1.9041	said
1.5876	reuters
1.4786	photo
1.4026	film
1.3945	also
1.2465	percent
1.2073	says

Bigrams

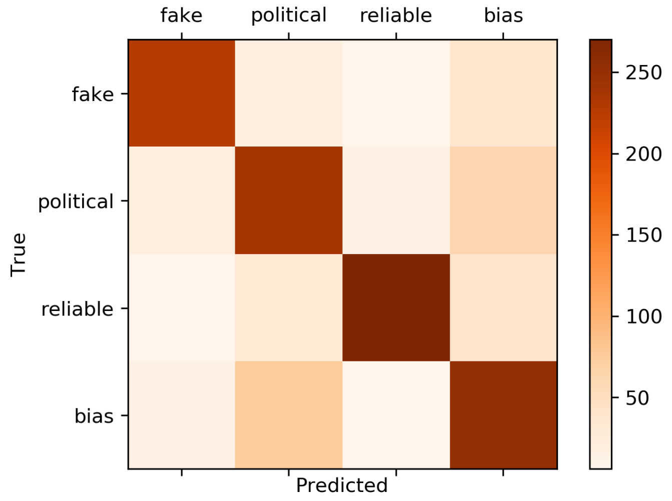
-4.2017	budget rep
-4.5269	aliens tend
-4.6422	aiding abetting
-4.6732	asking doctor
-4.7910	becoming nurse
-5.0097	books hillbilly
-5.2696	500 name
-5.3058	black sea
-5.3252	bar great
-5.3520	apartment metrocare
-5.3822	cabinet bloomberg
-5.3951	babies kinkade

Multiclass Classification with count vectorization

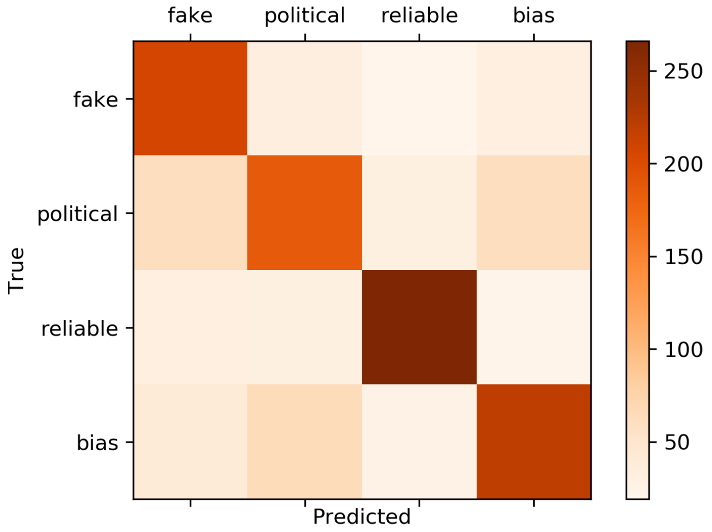
Confusion matrix of MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)



Confusion matrix of XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1, gamma=0, learning_rate=0.1, max_delta_step=0, max_depth=3, min_child_weight=1, missing=None, n_estimators=100, n_jobs=1, nthread=None, objective='multi:softprob', random_state=0, reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None, silent=None, subsample=1, verbosity=1)

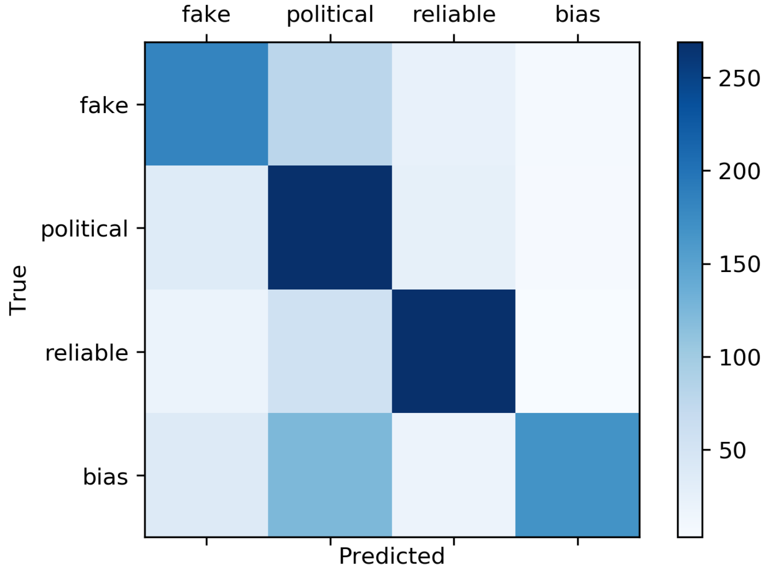


Confusion matrix of LinearSVC(C=1.0, class_weight=None, dual=True, fit_intercept=True, intercept_scaling=1, loss='squared_hinge', max_iter=1000, multi_class='ovr', penalty='l2', random_state=None, tol=0.0001, verbose=0)

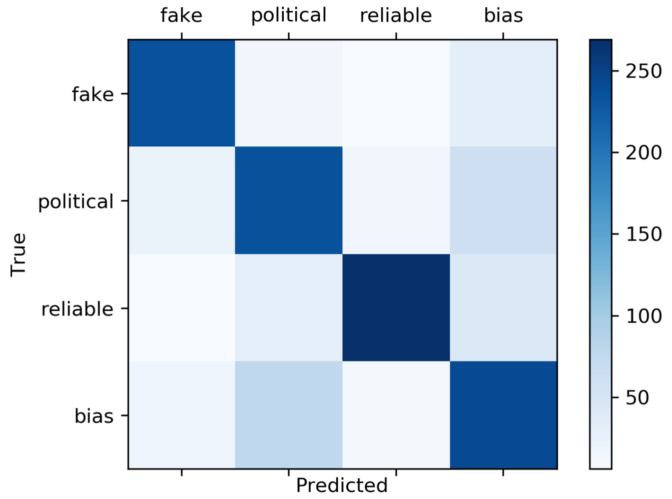


Multiclass Classification with tf-idf vectorization

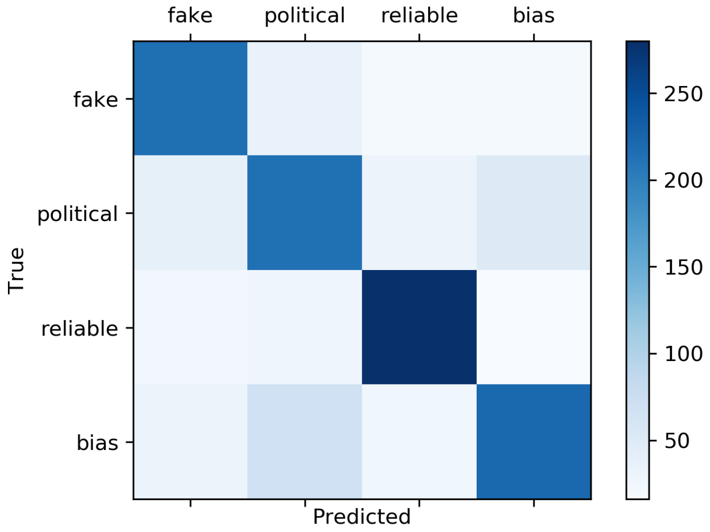
Confusion matrix of MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)



Confusion matrix of XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1, gamma=0, learning_rate=0.1, max_delta_step=0, max_depth=3, min_child_weight=1, missing=None, n_estimators=100, n_jobs=1, nthread=None, objective='multi:softprob', random_state=0, reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None, silent=None, subsample=1, verbosity=1)

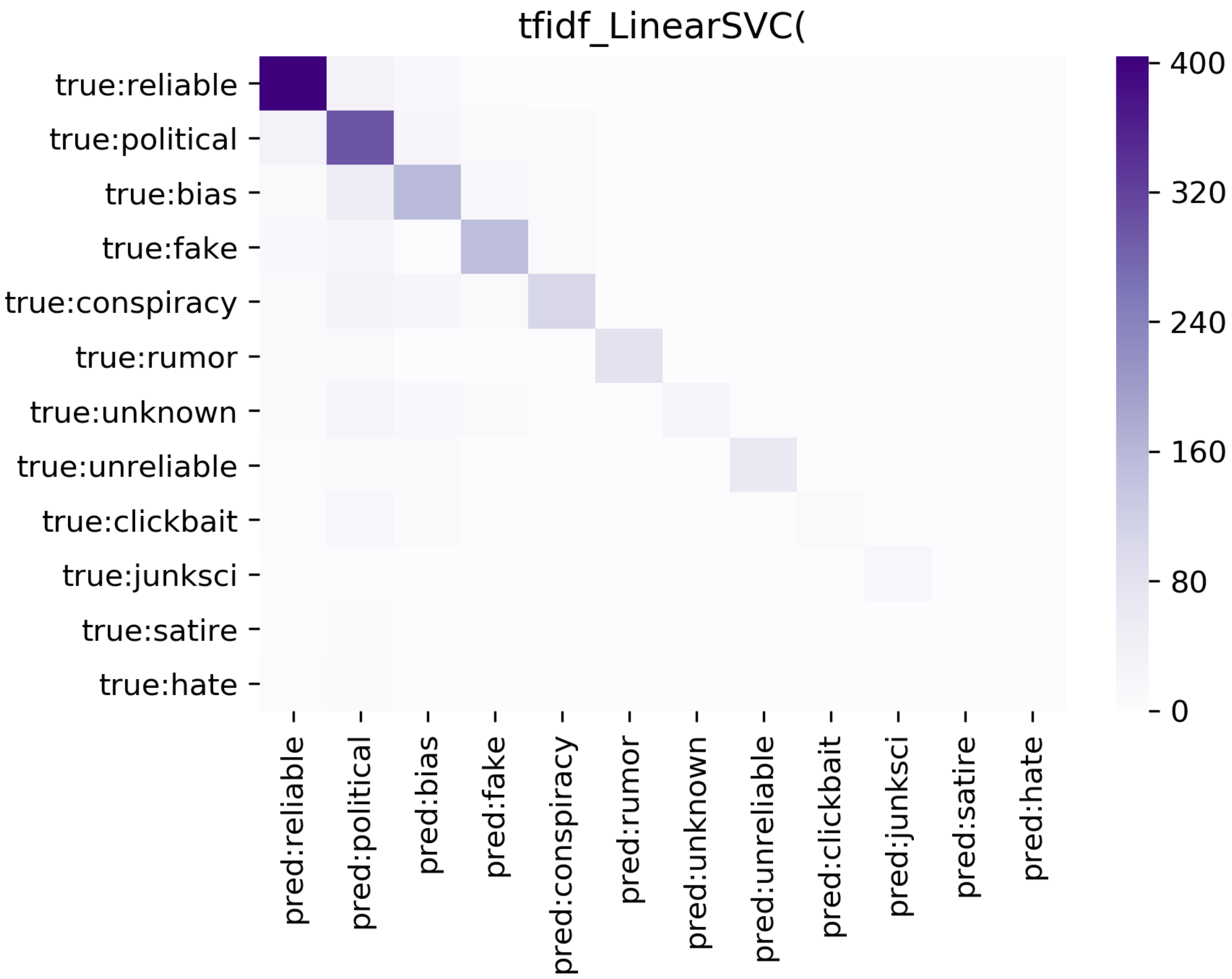


Confusion matrix of LinearSVC(C=1.0, class_weight=None, dual=True, fit_intercept=True, intercept_scaling=1, loss='squared_hinge', max_iter=1000, multi_class='ovr', penalty='l2', random_state=None, tol=0.0001, verbose=0)



More Classes

- The same six pairs of vectorization and classifiers were applied to the data, but for all of the classes (instead of just the four largest ones).



Concluding Points

- Prediction accuracy seems to be inherently tied to source.
- Some of the most informative features were direct references to the site or online newspaper that articles came from.
- This analysis was conducted with a smaller subset of data from the very large dataset.
- Conducting the analysis with a larger amount of data may yield more significant results.