Topic Classification
The New York Times Comments

Maria Annecchione
January 2020

Outline

Motivation
Examining the Dataset
Text Pre-processing
Feature Engineering
Feature Selection
Algorithm Evaluation
Ensemble Methods
Neural Networks
Interpretation

Why classify news comments?

Enhanced Reader Experience

Reader comments delve into **multiple topics**

Links to **related news item**

Previous work: **20 news-group dataset**

Examining the dataset

A kaggle dataset…

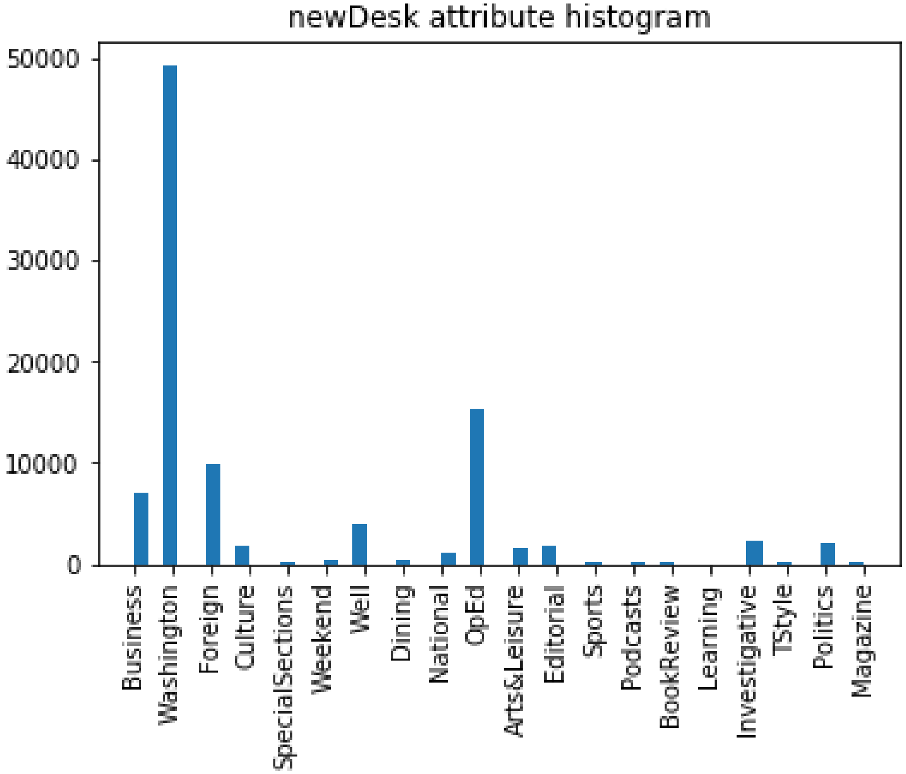
The March 2018 csv file

246946 x 33 data frame → 97308 comments

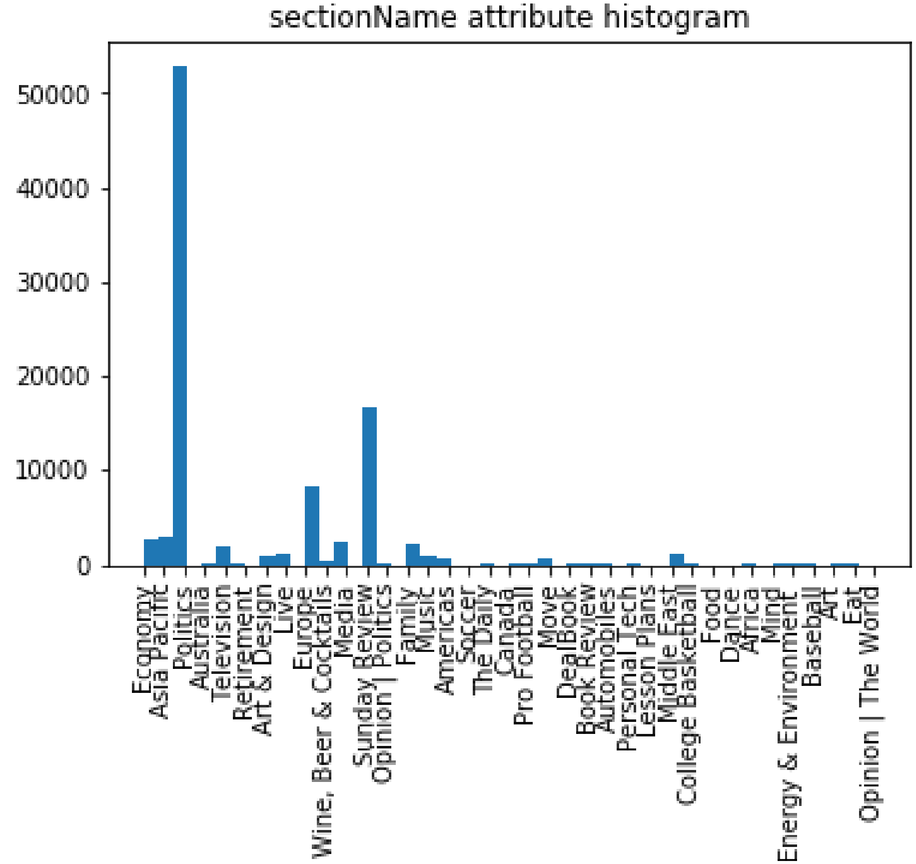
Columns of interest:

	commentBody	newDesk	sectionName
count	97308	97308	97308
unique	97179	20	37
top	Amen!	Washington	Politics
freq	5	49247	52798

Examining the dataset

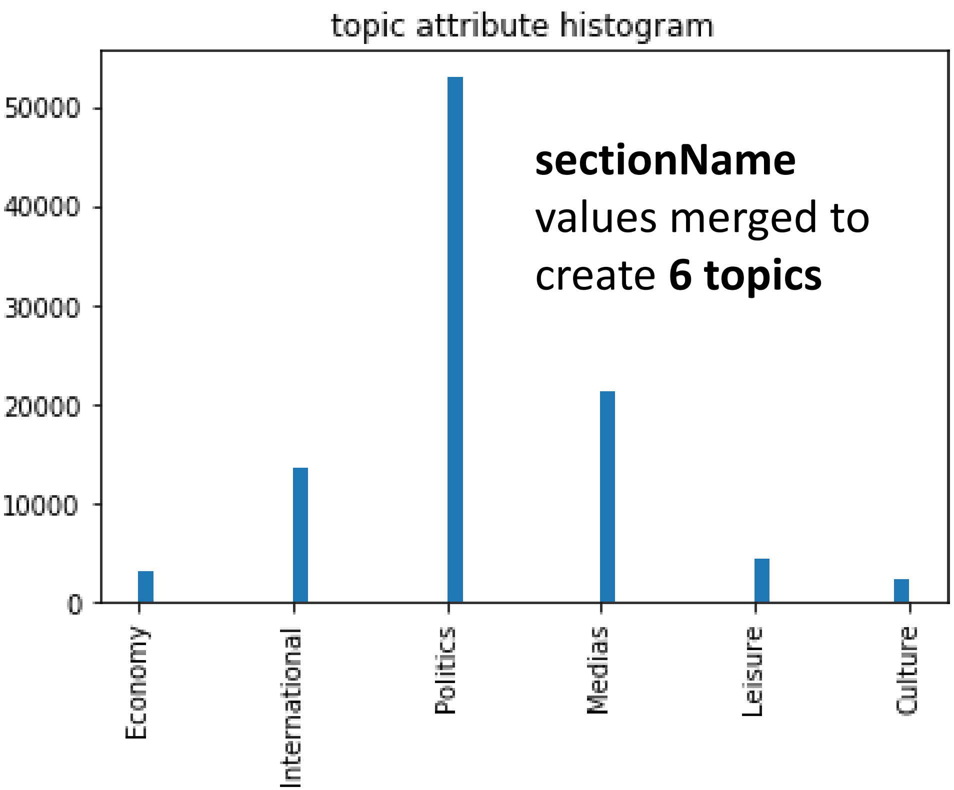


20 potential topics

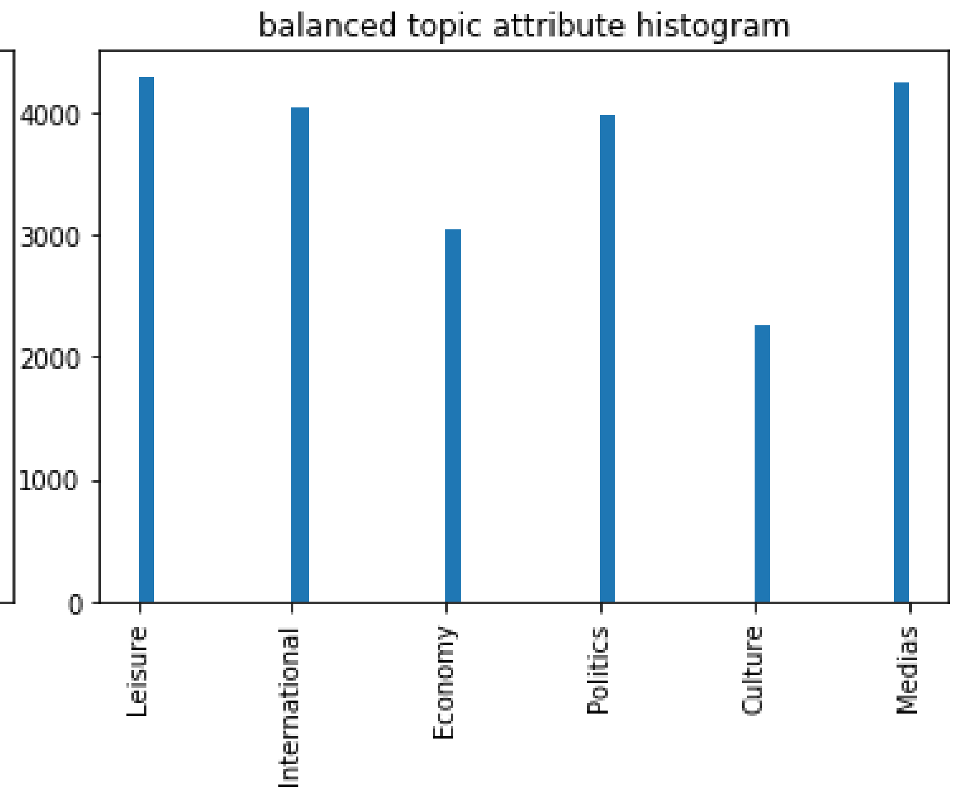


37 potential topics

The target variable…



sectionName values merged to create **6 topics**



	commentBody	topic
count	21844	21844
unique	21832	6
top	Thanks!	Leisure
freq	3	4292

Text pre-processing…

Before pre-processing:

Our beagles tend to howl or wake up barking when things go bump in the night, which can be super annoying at 3am. However, I have found turning on a fan or a bit of white noise helps them sleep through the night without getting woken (and thereby waking us) when they hear the train or a neighborhood cat.

After pre-processing:

beagle tend howl wake bark thing go bump night super annoy However find turn fan bite white noise help sleep night without get wake thereby wake us hear train neighborhood cat

Text pre-processing…

Before pre-processing:

Alabama has a $2 billion steel mill in Calvert, jointly owned by Japanese and German steel companies, that imports steel from Mexico and Brazil where it is fabricated for use in the many auto plants here in the a South. My question is has any in the White House come to grips that this is a perfect example of globalization?

After pre-processing:

Alabama billion steel mill Calvert jointly Japanese German steel company import steel Mexico Brazil fabricate use many auto plant South question White House come grip perfect example globalization

Text pre-processing…

Removing Empties…

There is at least one empty in pre-processed comments... being excluded.
Original comment: Why not 12?
There is at least one empty in pre-processed comments... being excluded.
Original comment: 25 him. Now.

Removing tokens that occur once…

50 most frequents words:

[('much', 8041), ('Trump', 7603), ('would', 4991), ('get', 4233), ('people', 4153), ('one', 3967), ('make', 3924), ('like', 3727), ('good', 3606), ('go', 3568), ('time', 3147), ('think', 3072), ('know', 3006), ('year', 2970), ('say', 2698), ('see', 2543), ('us', 2521), ('need', 2504), ('take', 2493), ('work', 2392), ('many', 2376), ('doe', 2340), ('US', 2309), ('little', 2271), ('want', 2242), ('country', 2158), ('may', 2142), ('way', 2099), ('even', 2071), ('American', 1986), ('use', 1932), ('well', 1888), ('live', 1863), ('world', 1861), ('thing', 1839), ('come', 1773), ('vote', 1643), ('right', 1612), ('give', 1609), ('long', 1592), ('man', 1552), ('day', 1544), ('also', 1506), ('child', 1470), ('Putin', 1438), ('never', 1432), ('find', 1399), ('job', 1368), ('Russia', 1342), ('great', 1312)]

total 20238 words

Feature Engineering

Four types of features:

- Bag-of-Words including uni- and bi-grams
- TF-IDF
- Word2Vec Model
- GloVe Model

Learning algorithm:

- Support Vector Machine

TF-IDF features…

Support Vector Machine

Model Performance metrics:

Accuracy: 0.6868

Precision: 0.6841

Recall: 0.6868

F1 Score: 0.6815

Model Classification report:

----

	precision	recall	f1-score	support
Economy	0.70	0.76	0.73	587
International	0.69	0.69	0.69	809
Politics	0.59	0.50	0.54	788
Medias	0.62	0.57	0.59	857
Leisure	0.73	0.91	0.81	862
Culture	0.85	0.70	0.77	462
accuracy			0.69	4365
macro avg	0 .70	0.69	0.69	4365
weighted avg	0.68	0.69	0.68	4365

Prediction Confusion Matrix:

Predicted:

	Economy	International	Politics	Medias	Leisure	Culture
Actual: Economy	448	32	38	37	28	4
International	52	558	81	57	56	5
Politics	77	121	393	132	52	13
Medias	46	66	126	489	108	22
Leisure	12	15	8	30	786	11
Culture	9	22	17	42	48	324

Word2Vec features…

Support Vector Machine
Model Performance metrics:

Accuracy: 0.5526
Precision: 0.5492
Recall: 0.5526
F1 Score: 0.5356

Model Classification report:
----

	precision	recall	f1-score	support
Economy	0.56	0.71	0.62	587
International	0.50	0.67	0.57	809
Politics	0.50	0.29	0.37	788
Medias	0.46	0.37	0.41	857
Leisure	0.65	0.83	0.73	862
Culture	0.69	0.41	0.51	462
accuracy			0.55	4365
macro avg	0.56	0.55	0.54	4365
weighted avg	0.55	0.55	0.54	4365

Prediction Confusion Matrix:

	Predicted: Economy	International	Politics	Medias	Leisure	Culture
Actual: Economy	414	56	34	34	40	9
International	78	546	63	64	48	10
Politics	113	252	232	134	42	15
Medias	74	173	118	318	137	37
Leisure	39	30	6	59	714	14
Culture	23	39	10	85	117	188

Feature Selection…

Four feature extraction methods on TF-IDF features:

- Univariate feature selection with F-Test for feature scoring
- Singular Value Decomposition
- Extra Trees Classifier
- Recursive Feature Elimination with Support Vector Machine

Classification accuracy without selecting features: **69%**

Number of features: **414 006**

Feature Selection…

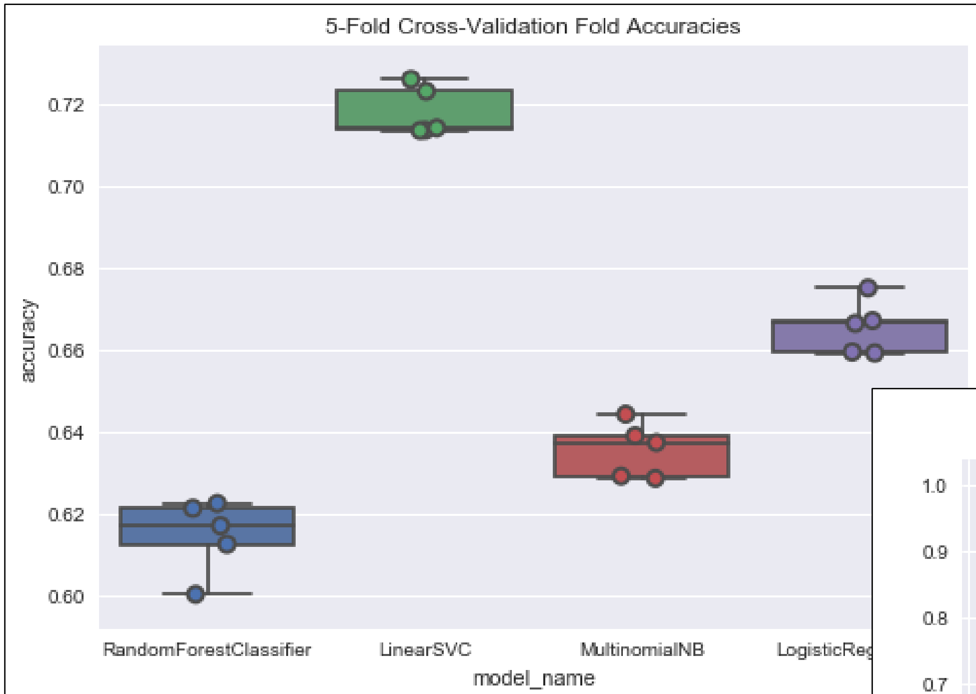
	Classification accuracy	Number of features
Univariate with F-test	69%	50000
Singular Value Decomposition	31%	500
Extra Trees Classifier	69%	51556
Recursive Feature Elimination (RFE)	69%	10000

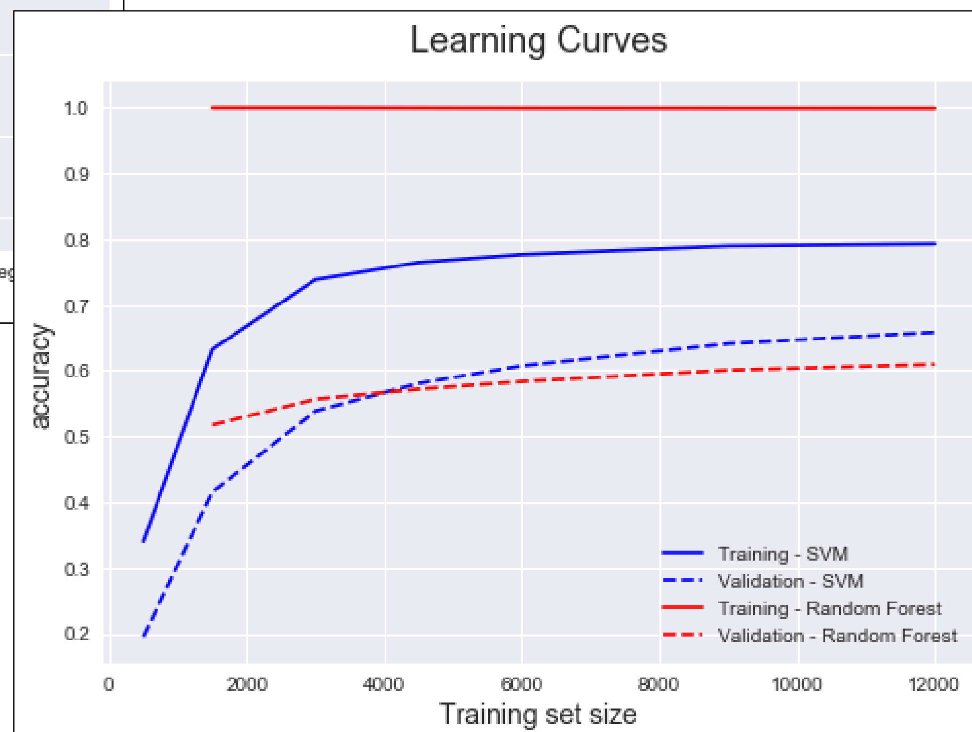
Algorithm Evaluation

Algorithms applied to RFE-selected TD-IDF features:

- Support Vector Machine
- Naive Bayes
- Logistic Regression
- Random Forest Classifier

Algorithm Evaluation





Algorithm Evaluation

SVM Performance Metrics and Classification Report

Model Performance metrics:

Accuracy: 0.6855
Precision: 0.6847
Recall: 0.6855
F1 Score: 0.6841

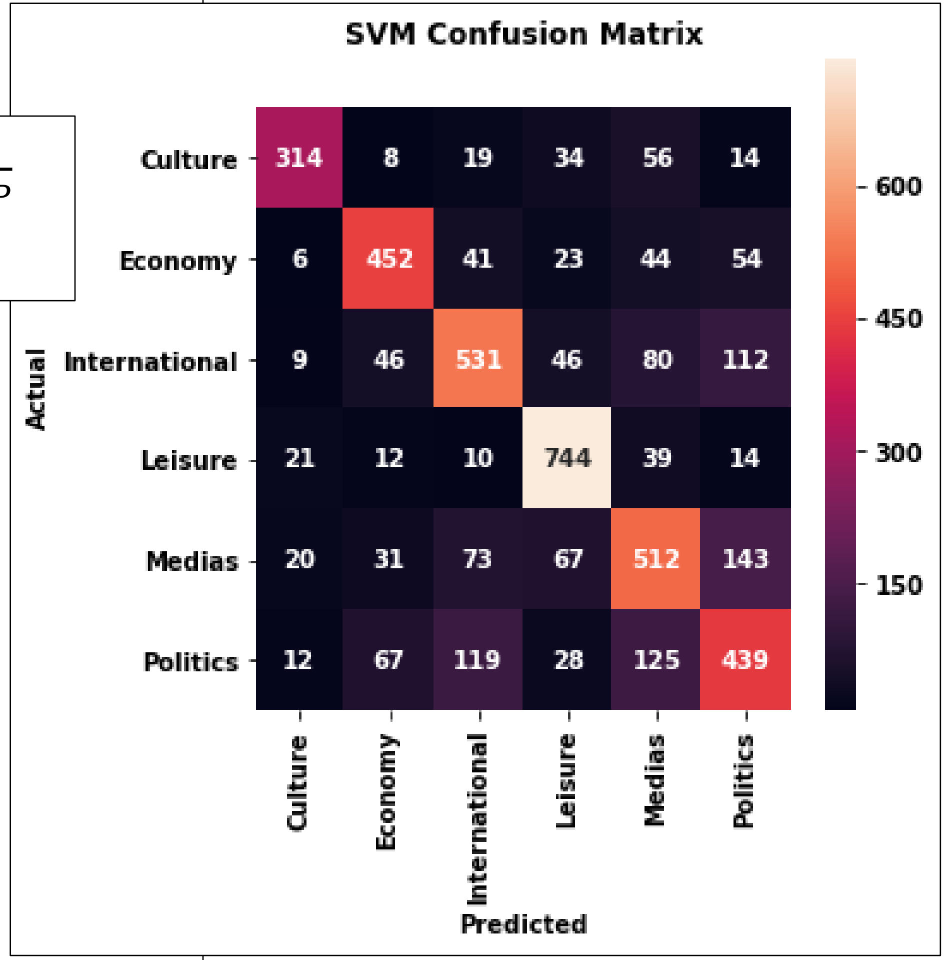
Precision : $\frac{TP}{TP+FP}$

Recall : $\frac{TP}{TP+FN}$

Model Classification report:

----

	precision	recall	f1-score	support
Economy	0.73	0.73	0.73	620
International	0.67	0.64	0.66	824
Politics	0.57	0.56	0.56	790
Medias	0.60	0.61	0.60	846
Leisure	0.79	0.89	0.84	840
Culture	0.82	0.71	0.76	445
accuracy			0.69	4365
macro avg	0.70	0.69	0.69	4365
weighted avg	0.68	0.69	0.68	4365



Algorithm Evaluation

Random Forest Performance Metrics and Classification Report

Model Performance metrics:

Accuracy: 0.6105
Precision: 0.6088
Recall: 0.6105
F1 Score: 0.6039

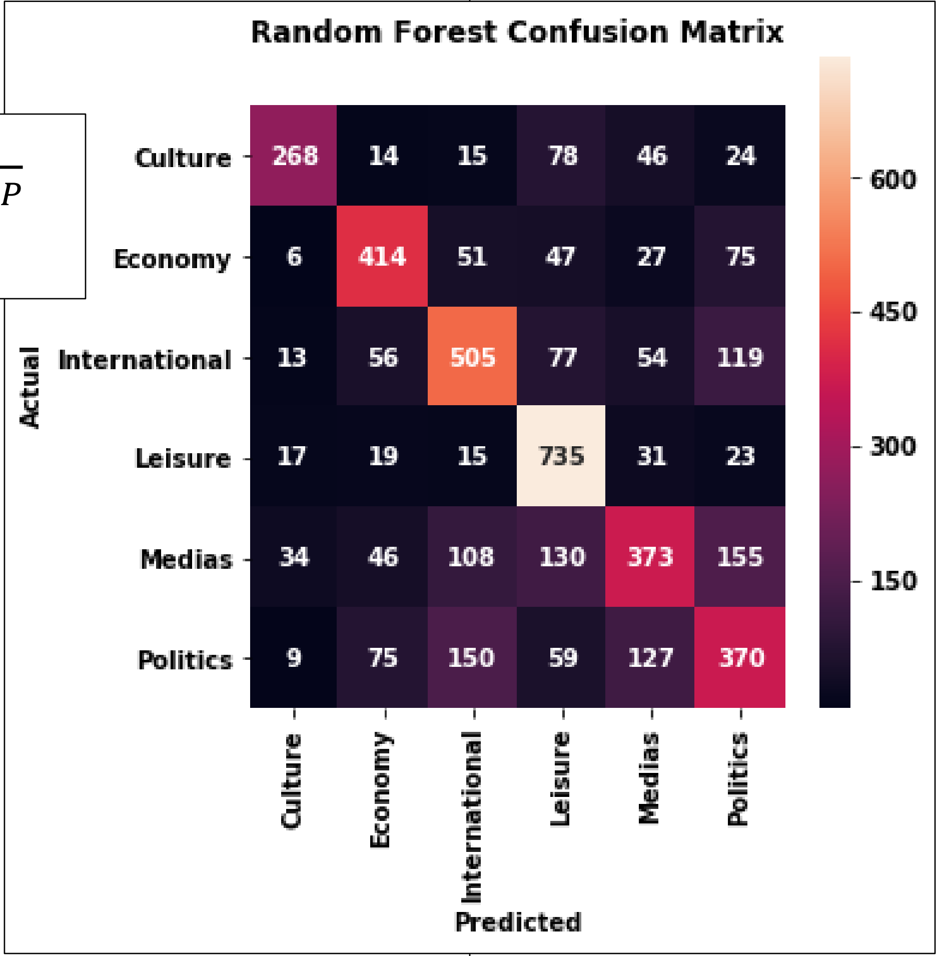
Precision : $\frac{TP}{TP+FP}$

Recall : $\frac{TP}{TP+FN}$

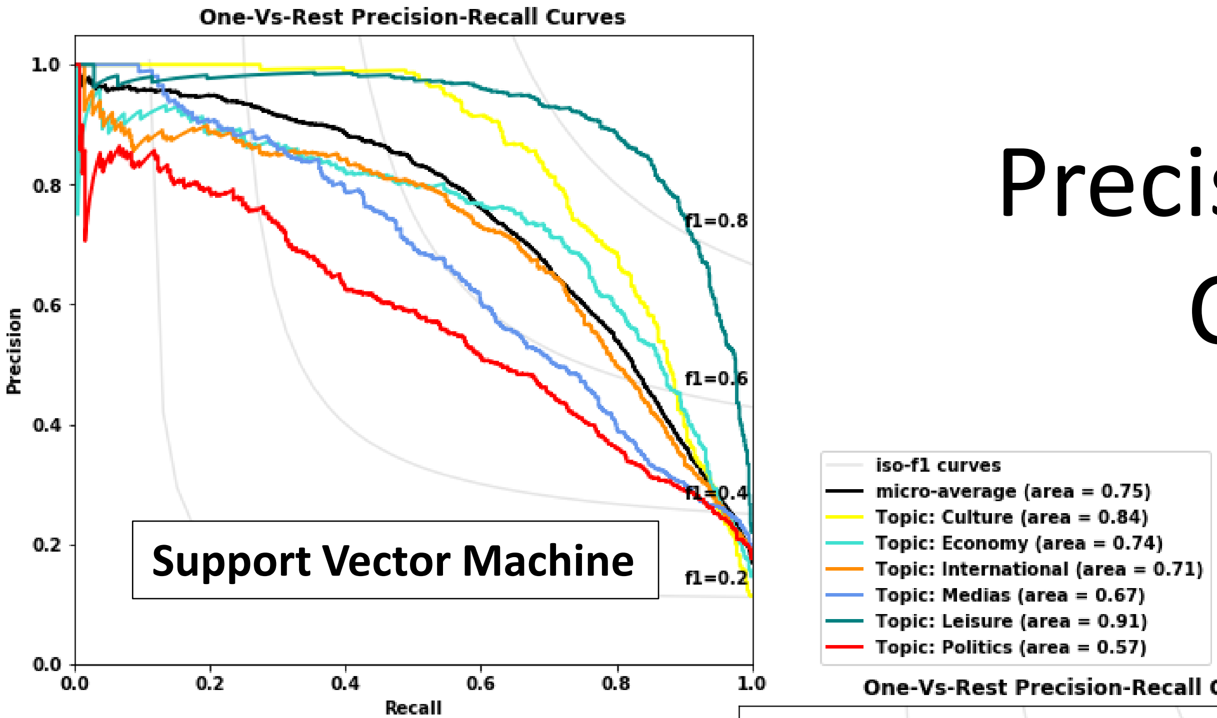
Model Classification report:

----

	precision	recall	f1-score	support
Economy	0.66	0.67	0.67	620
International	0.60	0.61	0.61	824
Politics	0.48	0.47	0.48	790
Medias	0.57	0.44	0.50	846
Leisure	0.65	0.88	0.75	840
Culture	0.77	0.60	0.68	445
accuracy			0.61	4365
macro avg	0.62	0.61	0.61	4365
weighted avg	0.61	0.61	0.60	4365

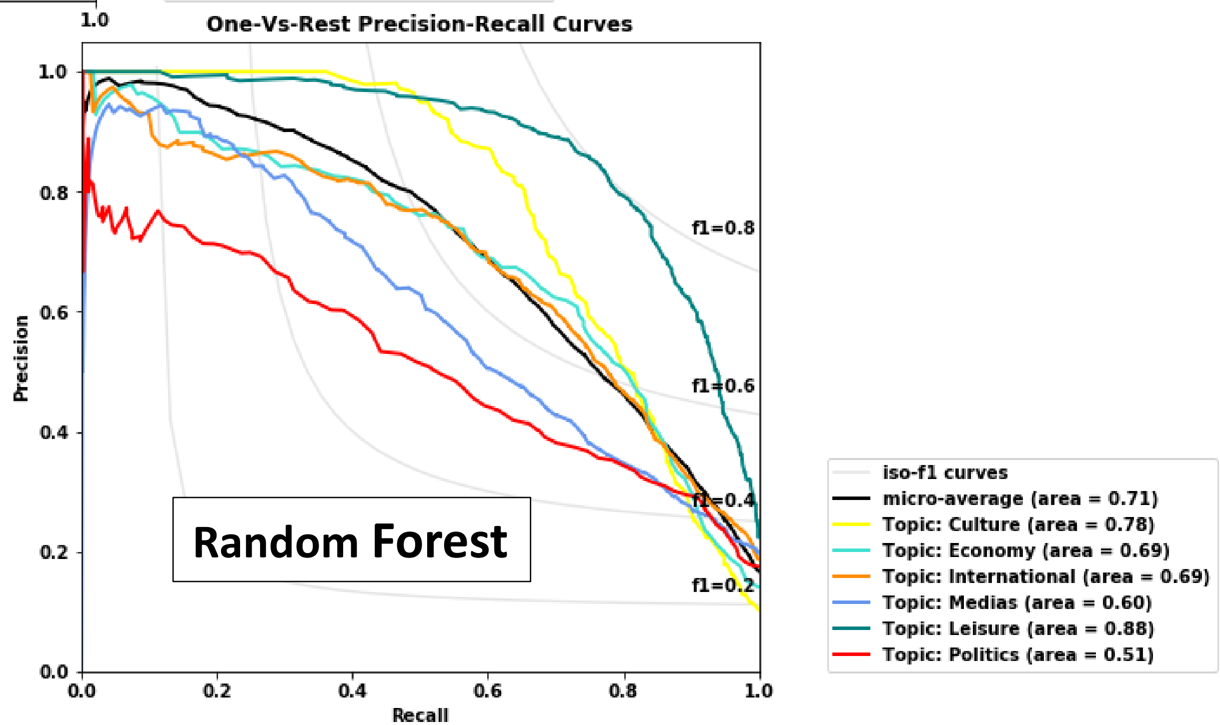


Precision-Recall Curves



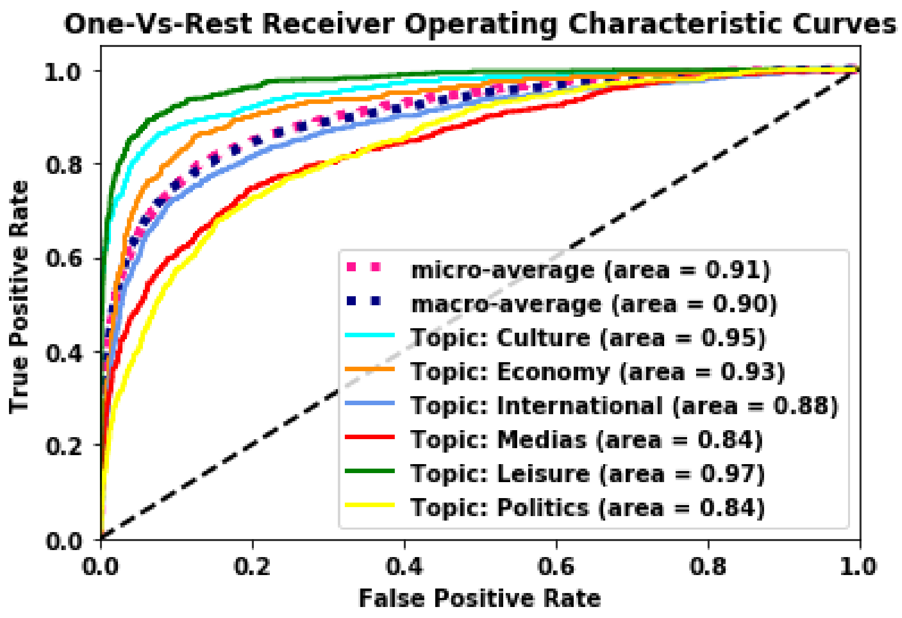
Precision : $\frac{TP}{TP+FP}$

Recall : $\frac{TP}{TP+FN}$

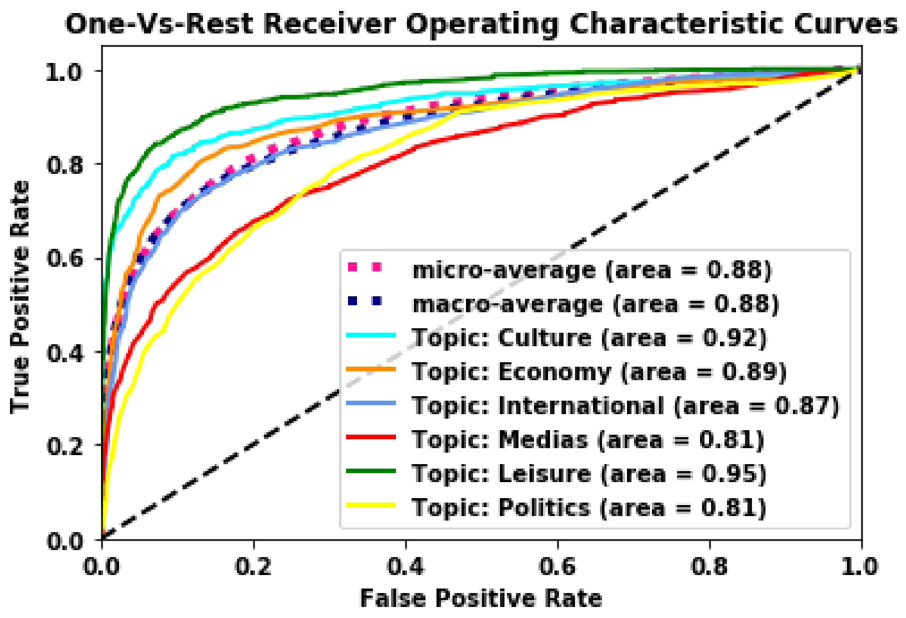


ROC curves

Support Vector Machine



Random Forest



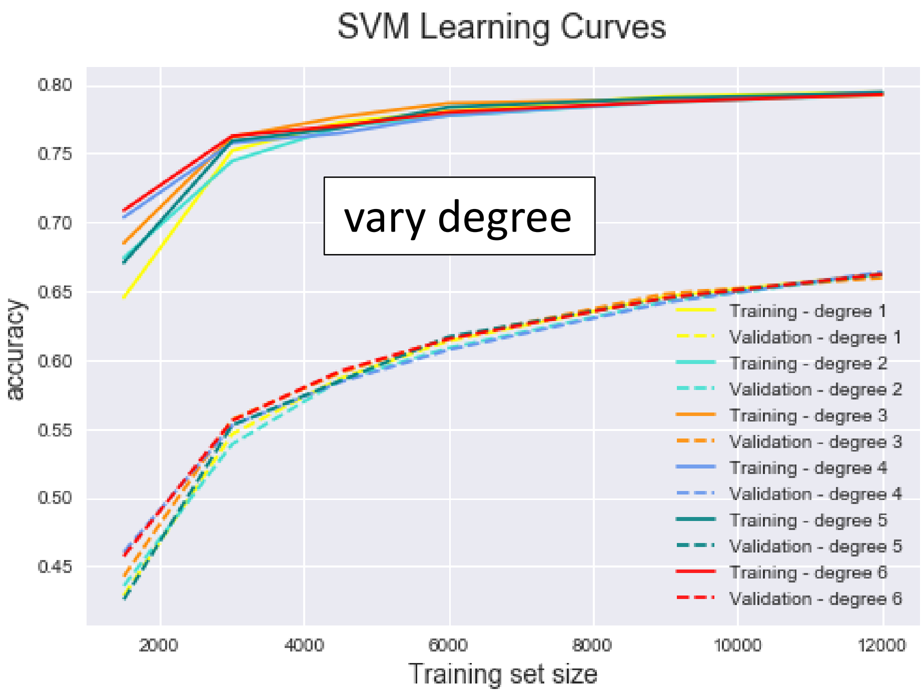
Precision : $\frac{TP}{TP+FP}$

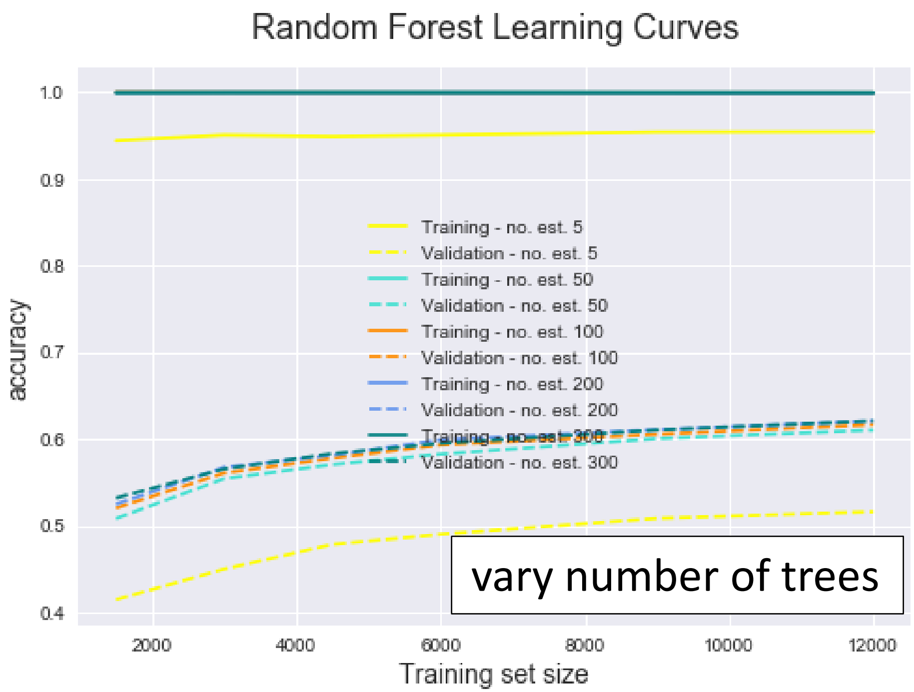
Recall : $\frac{TP}{TP+FN}$

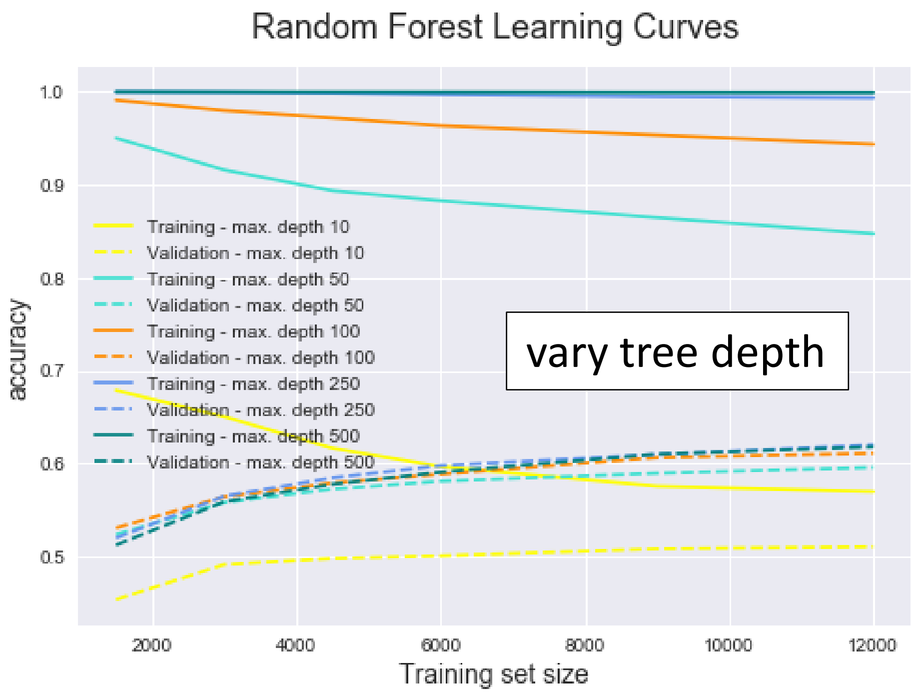
True Positive Rate: $\frac{TP}{TP+FP}$

False Positive Rate: $\frac{FP}{FP+TN}$

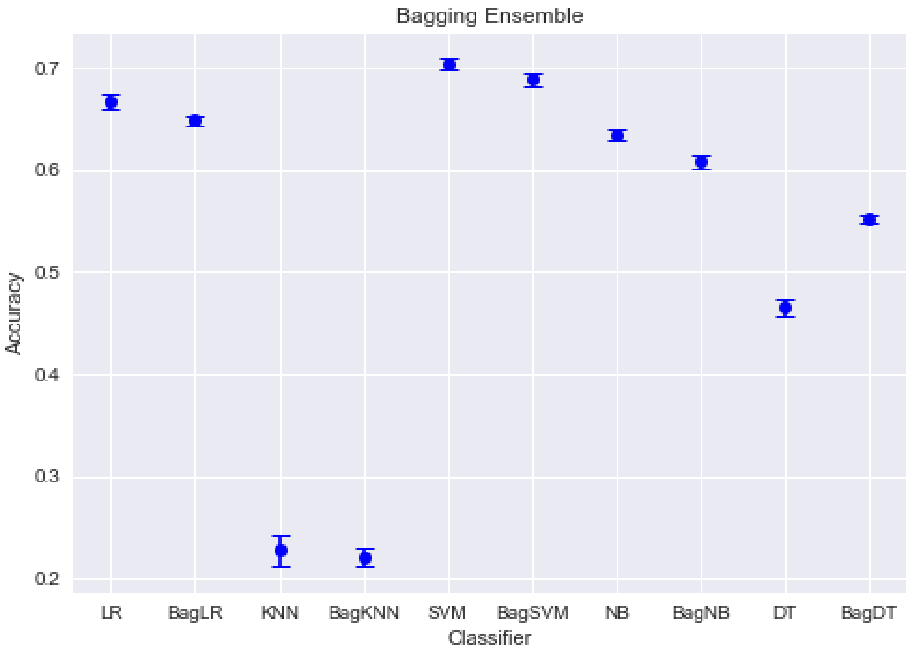
Hyper-parameter tuning

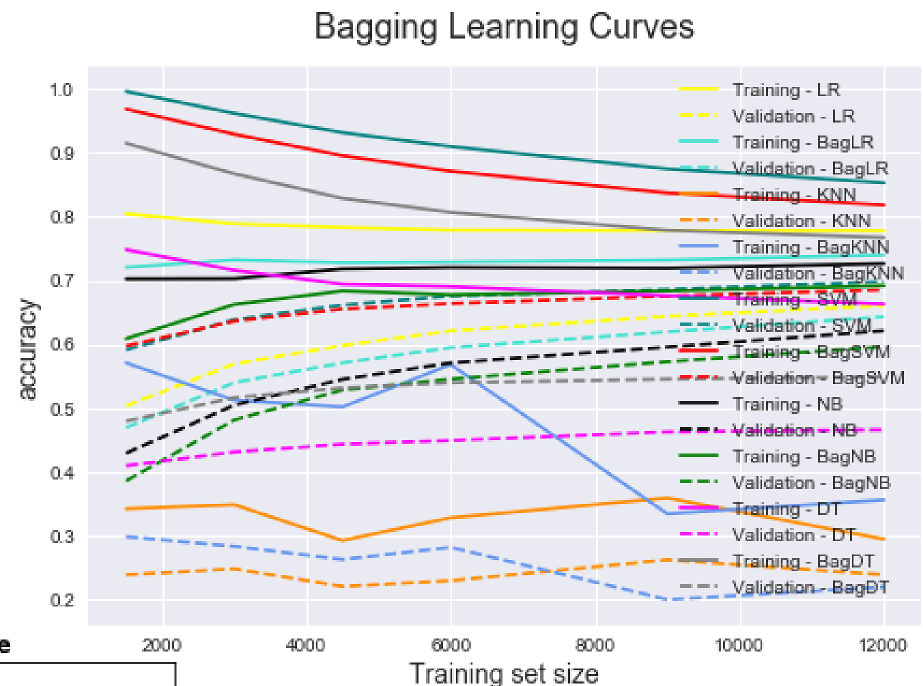


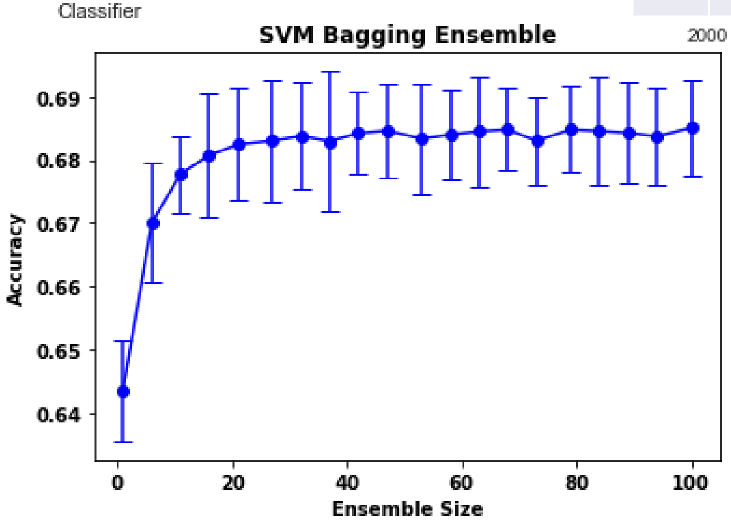




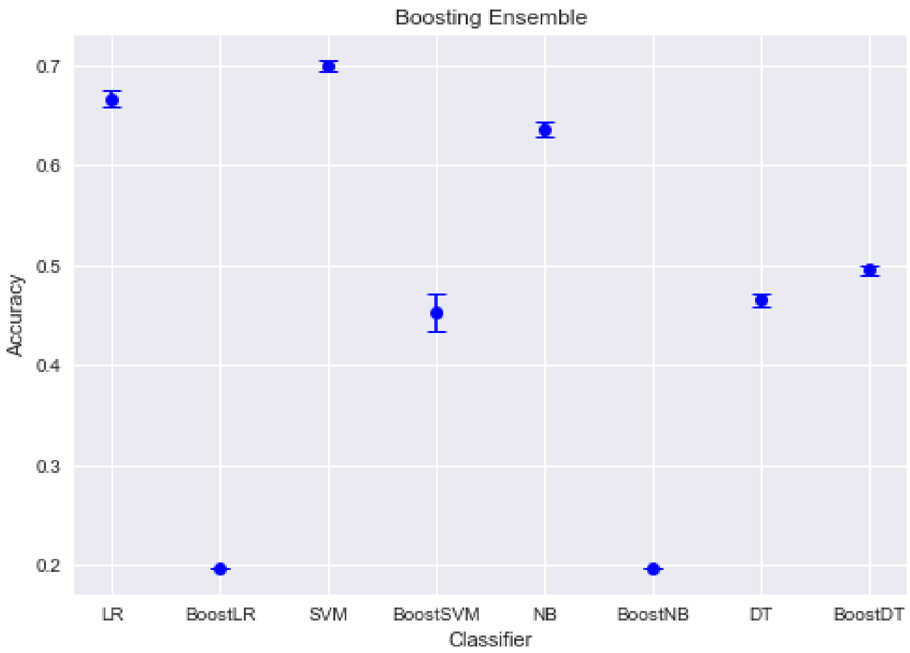
Ensemble Methods - Bagging

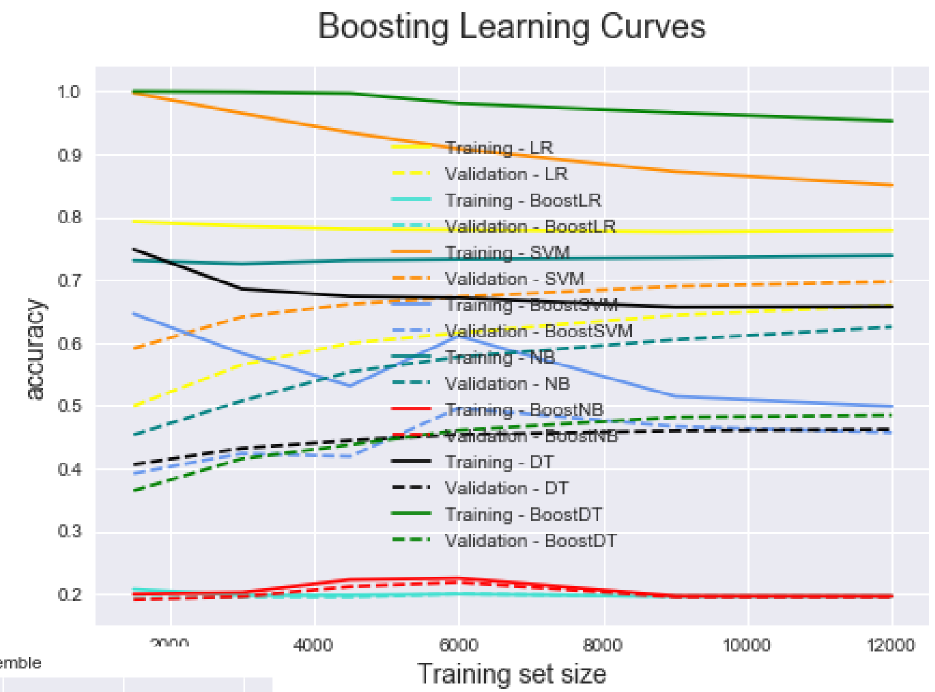




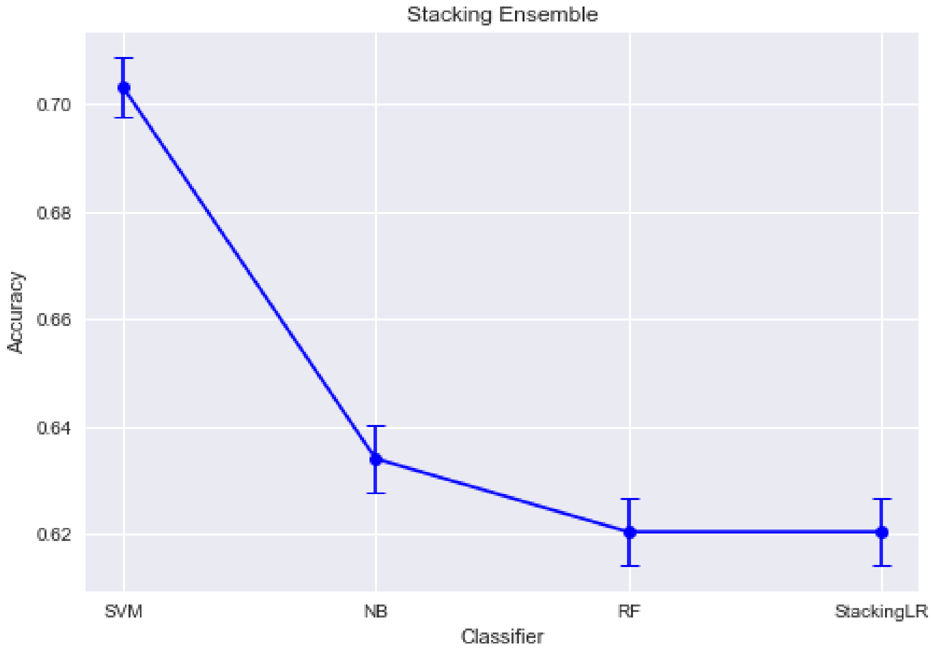


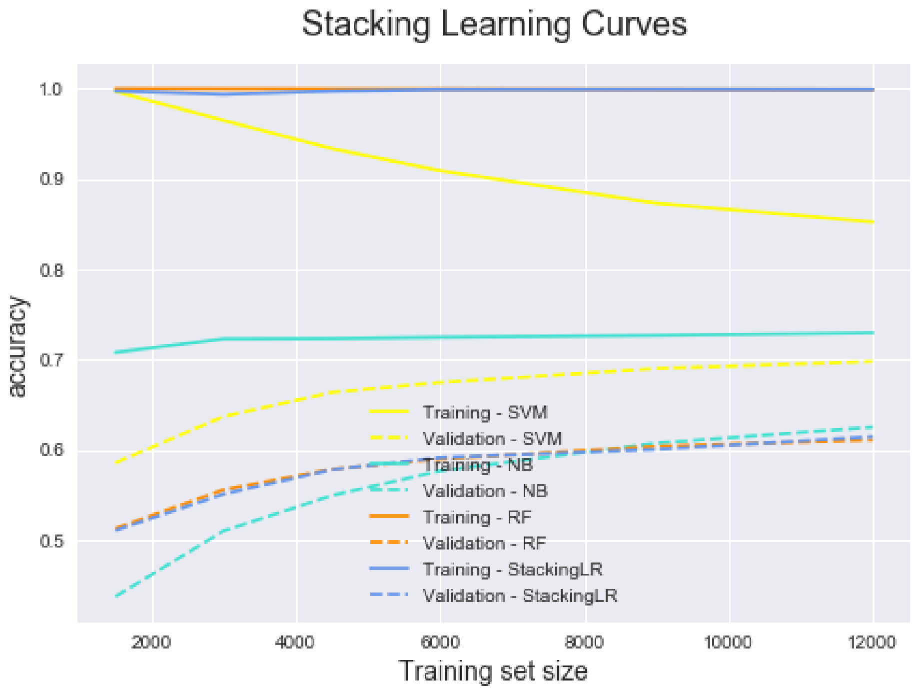
Ensemble Methods - Boosting





Ensemble Methods - Stacking

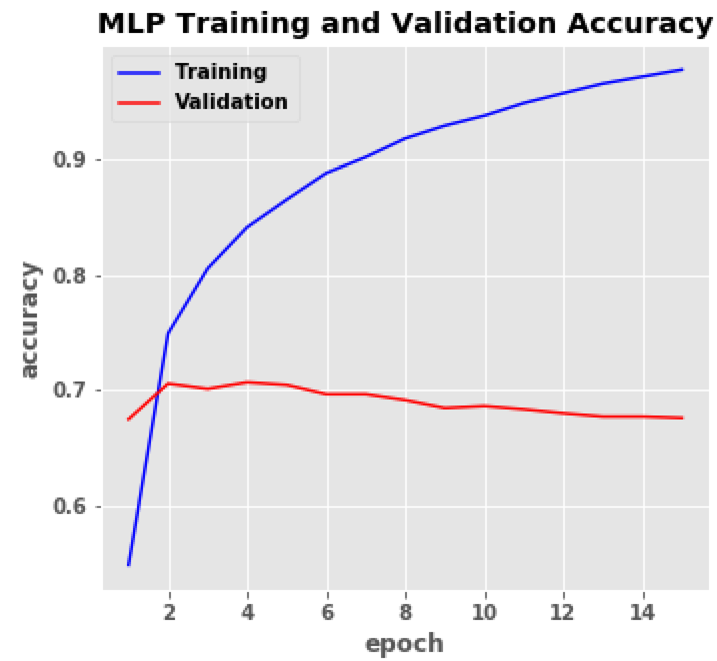




Neural Networks

Multi-Layer Perceptron

```
Layer (type)              Output Shape          Param #
=================================================================
dense_7 (Dense)           (None, 512)           2560512
_________________________________________________________________
activation_5 (Activation) (None, 512)           0
_________________________________________________________________
dropout_3 (Dropout)       (None, 512)           0
_________________________________________________________________
dense_8 (Dense)           (None, 6)             3078
_________________________________________________________________
activation_6 (Activation) (None, 6)             0
=================================================================
Total params: 2,563,590
Trainable params: 2,563,590
Non-trainable params: 0
_________________________________________________________________
Train on 15713 samples, validate on 1746 samples
Epoch 1/15
15713/15713 [==============================] - 21s 1ms/step - loss: 1.3440 - acc: 0.5488 - val_loss: 0.9832 - val_acc: 0.6747
.
.
.
Epoch 15/15
15713/15713 [==============================] - 20s 1ms/step - loss: 0.1034 - acc: 0.9775 - val_loss: 1.2880 - val_acc: 0.6758
```



Neural Networks

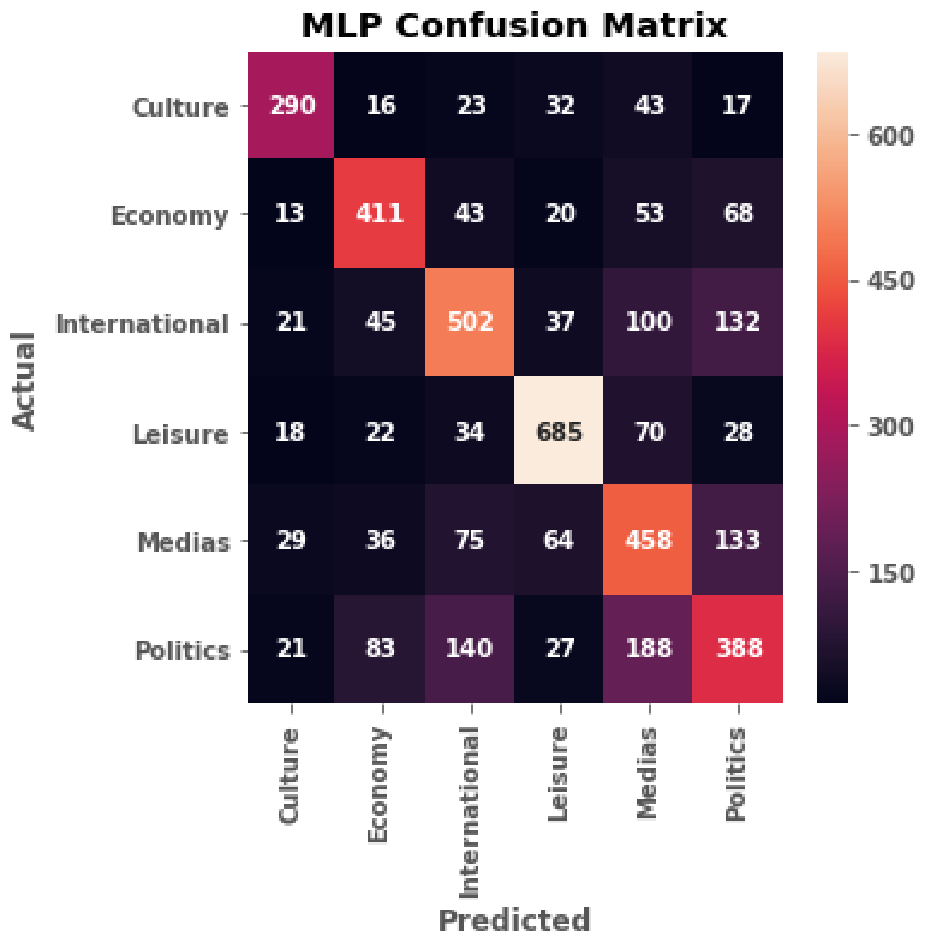
MLP Performance Metrics and Classification Report

Model Performance metrics:

Accuracy: 0.6263
Precision: 0.6278
Recall: 0.6263
F1 Score: 0.6263

Model Classification report:
----

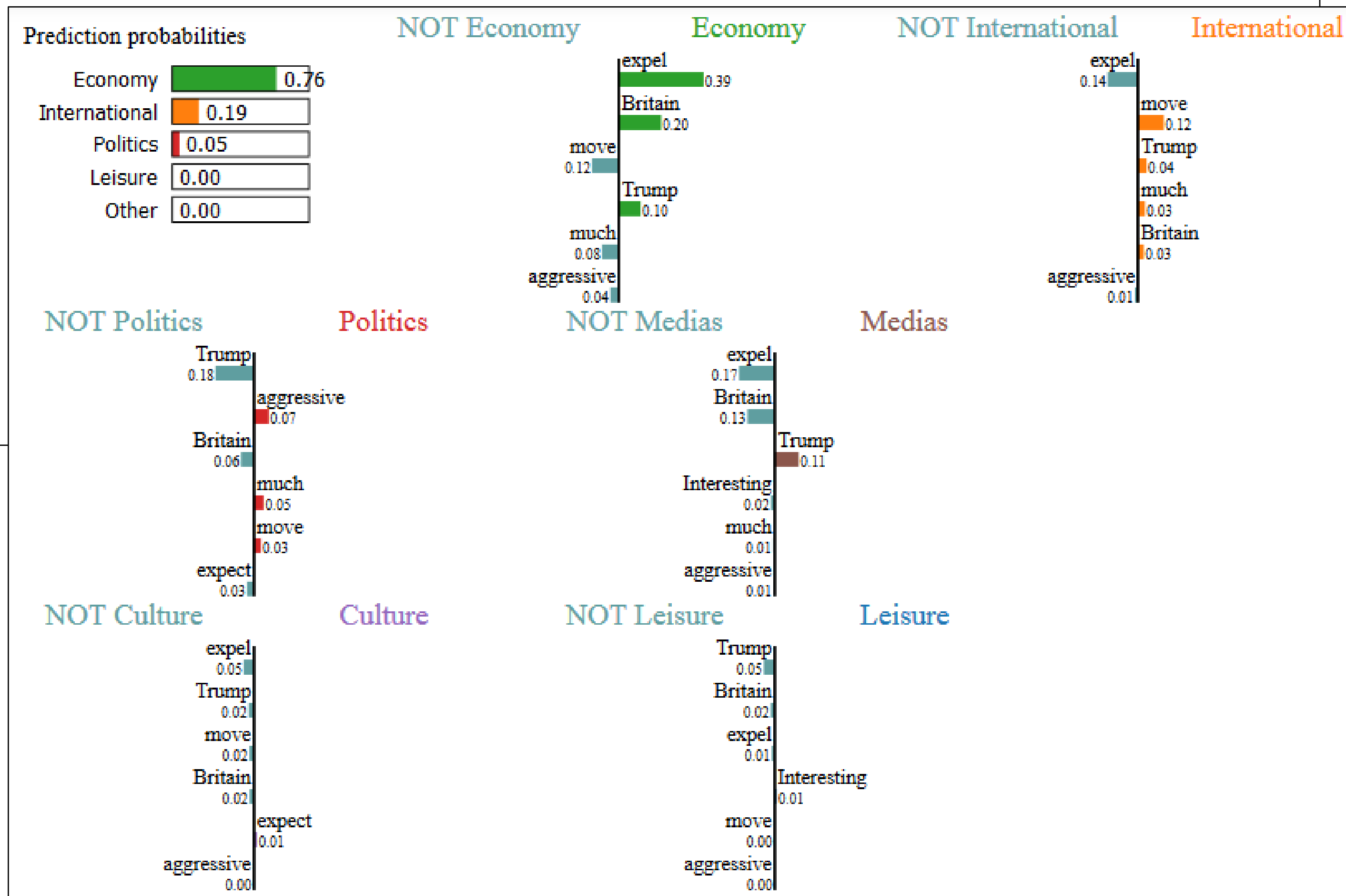
	precision	recall	f1-score	support
International	0.61	0.60	0.61	837
Politics	0.51	0.46	0.48	847
Economy	0.67	0.68	0.67	608
Medias	0.50	0.58	0.54	795
Leisure	0.79	0.80	0.80	857
Culture	0.74	0.69	0.71	421
accuracy			0.63	4365
macro avg	0.64	0.63	0.63	4365
weighted avg	0.63	0.63	0.63	4365



Interpretation - LIME

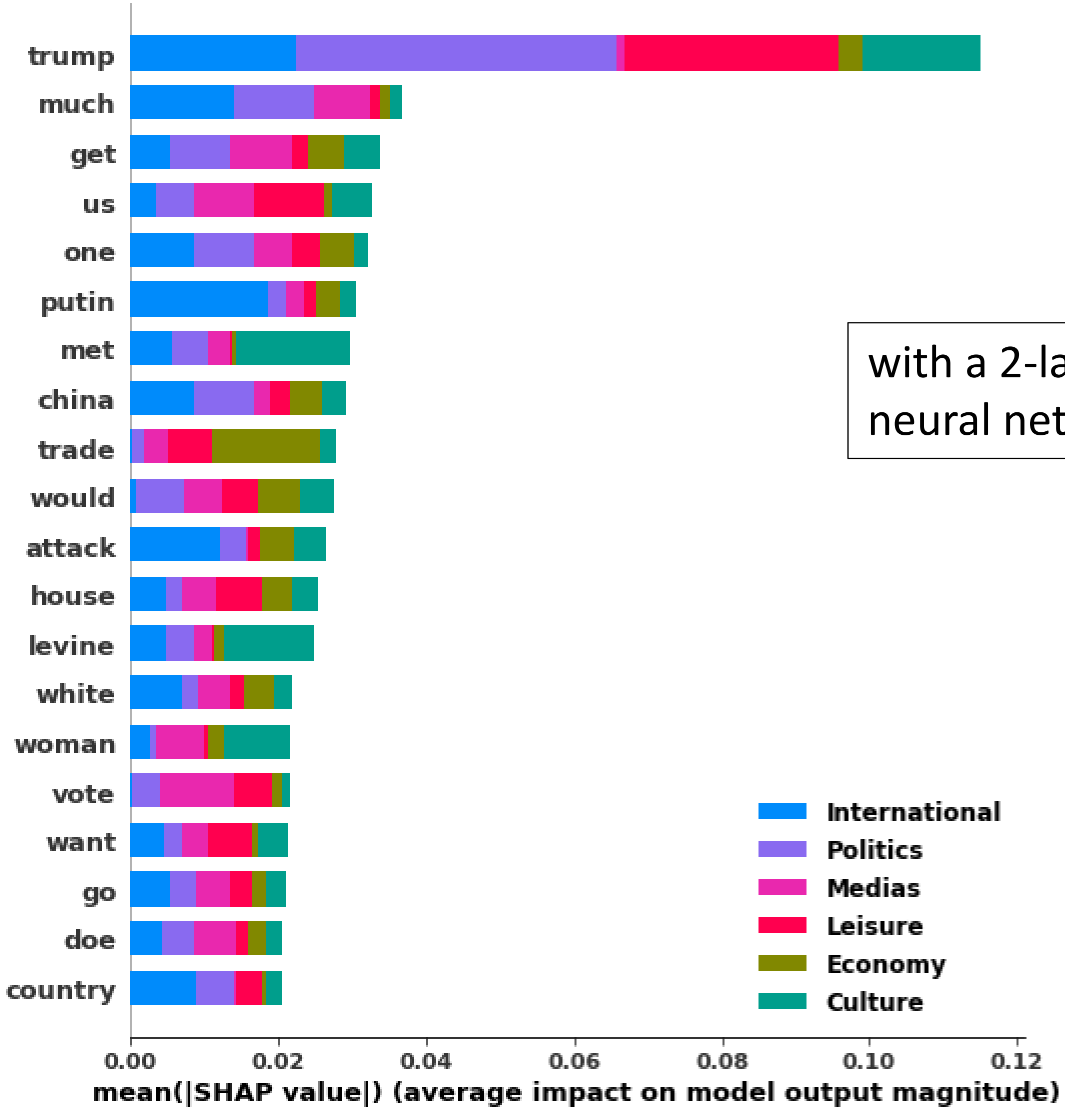
Comment id: 3818
Comment (cleaned): much aggressive move Trump expect Britain expel expel Interesting
Predicted Topic (SVM): International
True Topic: Economy

LIME ordering:
2 Economy
1 International
3 Politics
5 Medias
4 Culture
0 Leisure



Locally Interpretable Model-Agnostic Explanations

Interpretation - SHAP



with a 2-layer neural network

Shapley Additive Explanations

Conclusions

F1-score	Multilayer Perceptron	Random Forest	Support Vector Machine
International	60%	61%	66%
Politics	48%	48%	56%
Economy	67%	67%	73%
Medias	54%	50%	60%
Leisure	80%	75%	84%
Culture	71%	68%	76%

- **Leisure** topic is predicted most accurately
- **Politics** topic is predicted least accurately
- **Support Vector Machine** did best
- **TF-IDF features** did best

Future work…

- Try unsupervised learning methods
- Try features extracted from language models…word context is preserved

Thanks

References

Llewellyn, C., Grover, C. and Oberlander, J. (2016) **Improving Topic Model Clustering of Newspaper Comments for Summarisation**. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop, Berlin, Germany

Shaikh, Javed (2017) **Machine Learning, NLP: Text Classification using scikit-learn, python and NLTK**. Towards Data Science