Fake News Predictor

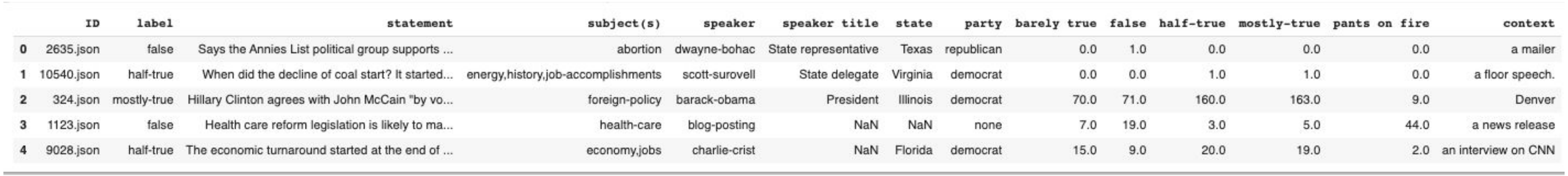
Nishant Date
Nick Tran

Table of Contents

- Dataset
- Data Preprocessing
- Logistic Regression
  - L1 and L2 Regularization
  - Polynomial Fit Transformation
- Support Vector Machines
  - Linear
  - RBF
  - Polynomial
- Neural Network
- Conclusion

Dataset

- The dataset used is Liar: A Benchmark Dataset for Fake News Detection from William Yang Wang
    - There are 14 features
    - 20,480 total examples



	ID	label	statement	subject(s)	speaker	speaker title	state	party	barely true	false	half-true	mostly-true	pants on fire	context
0	2635.json	false	Says the Annies List political group supports ...	abortion	dwayne-bohac	State representative	Texas	republican	0.0	1.0	0.0	0.0	0.0	a mailer
1	10540.json	half-true	When did the decline of coal start? It started...	energy,history,job-accomplishments	scott-surovell	State delegate	Virginia	democrat	0.0	0.0	1.0	1.0	0.0	a floor speech.
2	324.json	mostly-true	Hillary Clinton agrees with John McCain "by vo...	foreign-policy	barack-obama	President	Illinois	democrat	70.0	71.0	160.0	163.0	9.0	Denver
3	1123.json	false	Health care reform legislation is likely to ma...	health-care	blog-posting	NaN	NaN	none	7.0	19.0	3.0	5.0	44.0	a news release
4	9028.json	half-true	The economic turnaround started at the end of ...	economy,jobs	charlie-crist	NaN	Florida	democrat	15.0	9.0	20.0	19.0	2.0	an interview on CNN

Data Preprocessing

- Binary Classification of Label column
- Feature Encoding of State, Party and Label
- Polynomial Degree 3 Transformation

1	2	3	4	5	6	7	8	9
Label	State	Party	Barely True	False	half-true	mostly-true	Pants-on fire	Truth-score
Binary and Label Encoder	LabelEncoder	LabelEncoder	-	-	-	-	-	See below

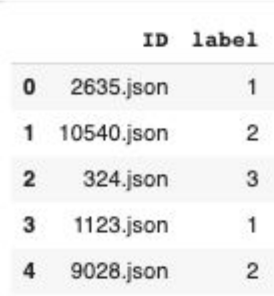
- Label = {'pants-fire':0, 'false':0,'barely-true':1, 'half-true':1, 'mostly-true':1, 'true':1}
- Truth-score = (barely true + half-true + mostly-true) / (total number of statements)

Logistic Regression

Table 2: Results for Binary-Classified Labels

Test	Regularization	Feature Transformation	Features Passed In	Training Accuracy	Testing Accuracy
1	L1	None	Features[2, 3, 4-8]	72.32%	73.1%
2	L2	None	Features[2, 3, 4-8]	73.80%	72.45%
3	L1	None	Features[9]	80.30%	80.43%
4	L2	None	Features[9]	80.30%	80.43%
5	L1	None	Features[2, 3, 9]	79.96%	80.82%
6	L2	None	Features[2, 3, 9]	80.24%	80.66%
7	L1	Polynomial Degree 3	Features[2, 3, 9]	80.00%	81.16%
8*	**L2**	**Polynomial Degree 3**	**Features[2, 3, 9]**	**79.96%**	**81.77%**
9	L1	Polynomial Degree 3	Features[2,3,4-9]	73.46%	74.66%
10	L2	Polynomial Degree 3	Features[2,3,4-9]	74.05%	76.00%

Logistic Regression with Non-binary classified Labels



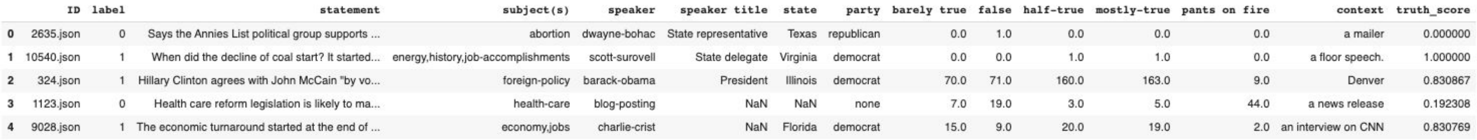
	ID	label
0	2635.json	1
1	10540.json	2
2	324.json	3
3	1123.json	1
4	9028.json	2

Test	Regularization	Feature Transformation	Features Passed In	Training Accuracy	Testing Accuracy
A	L1	None	Features[4-8]	32.67%	31.57%
B	L2	None	Features[4-8]	28.05%	27.70%
C	**L1**	**Polynomial Degree 3**	**Features[2,3,4-9]**	**34.46%**	**36.23%**
D*	L2	Polynomial Degree 3	Features[2,3,4-9]	31.96%	32.99%

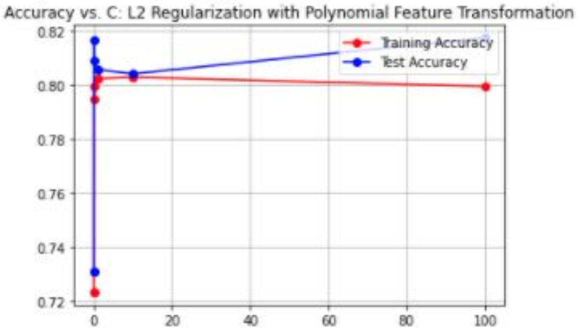
***Test C yielded the best results**

Logistic Regression (best test)

Regularization	Transformation	Features	Training Accuracy	Testing Accuracy
L2	Polynomial Degree 3	Features[2, 3, 9]	79.96%	81.77%



	ID	label	statement	subject(s)	speaker	speaker title	state	party	barely true	false	half-true	mostly-true	pants on fire	context	truth_score
0	2635.json	0	Says the Annies List political group supports ...	abortion	dwayne-bohac	State representative	Texas	republican	0.0	1.0	0.0	0.0	0.0	a mailer	0.000000
1	10540.json	1	When did the decline of coal start? It started...	energy,history,job-accomplishments	scott-surovell	State delegate	Virginia	democrat	0.0	0.0	1.0	1.0	0.0	a floor speech.	1.000000
2	324.json	1	Hillary Clinton agrees with John McCain "by vo...	foreign-policy	barack-obama	President	Illinois	democrat	70.0	71.0	160.0	163.0	9.0	Denver	0.830867
3	1123.json	0	Health care reform legislation is likely to ma...	health-care	blog-posting	NaN	NaN	none	7.0	19.0	3.0	5.0	44.0	a news release	0.192308
4	9028.json	1	The economic turnaround started at the end of ...	economy,jobs	charlie-crist	NaN	Florida	democrat	15.0	9.0	20.0	19.0	2.0	an interview on CNN	0.830769

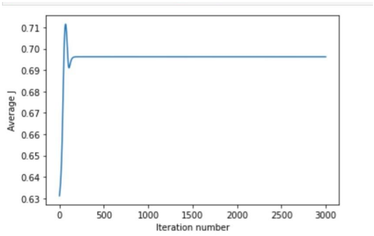
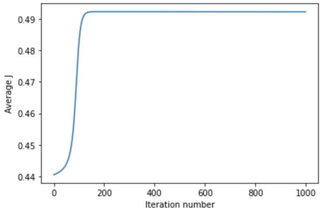


SVM

- Binary classified labels

Kernel	Training Accuracy	Training Accuracy
Linear	79.76%	81.37%
RBF	72.32%	73.08%
Polynomial	72.32%	73.08%

Neural Network

Hidden Layers - 5 Input Layers - 3 Output layers - 2 3000 iterations	 `Prediction accuracy is 56.27466456195738%` **56% accuracy**
Neural Network 1000 iterations 1 Output Layer 6 Hidden Layers	 43.6% Accuracy

Fake News Predictor

Nishant Date
Nick Tran

https://docs.google.com/presentation/d/14kvL_ppbMTA9wY1D_1r7xEGRY2qnv6JDRJRrt0aH0H8/edit#slide=id.gd5164d3991_0_10