Predicting the Source of News Articles Using NLP

Nicholas Thomas

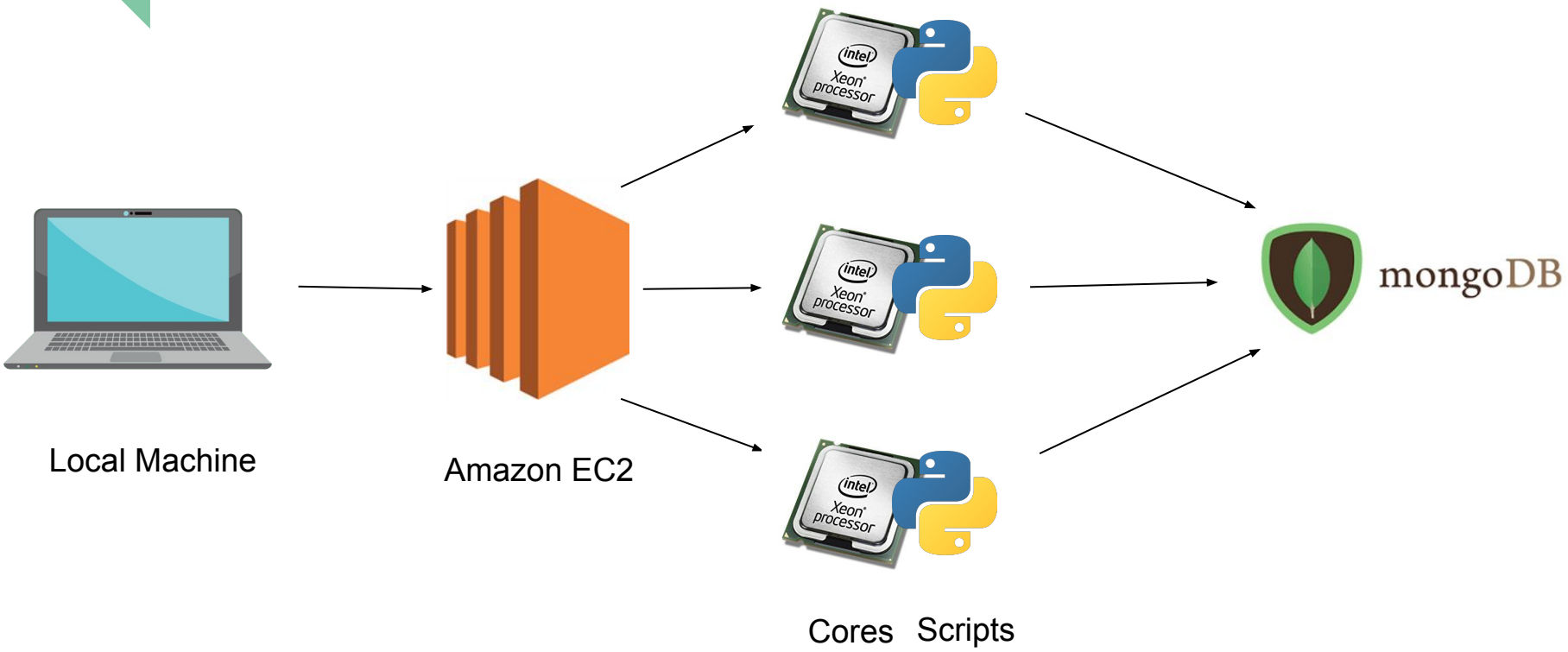
Objective

How can we predict an article’s source?

Sources

- 4092 political news articles
- 11 sources representing left, right, center, and conspiracy/fake

Process



Word stemming and stop words

They are a group of 13 highly partisan Democrats that make up the Mueller team, excluding him, (who) are trying very, very hard to frame him, to get him in trouble, when he hasn’t done anything wrong

Part of speech tagging

they ~~are a~~ group ~~of~~ 13 highli partisan democrat ~~that~~ make ~~up the~~ mueller team, exclud him, (who) ~~are~~ tri veri, veri hard ~~to~~ frame him, ~~to~~ get ~~him in~~ trouble, ~~when he~~ hasn’t done anyth wrong

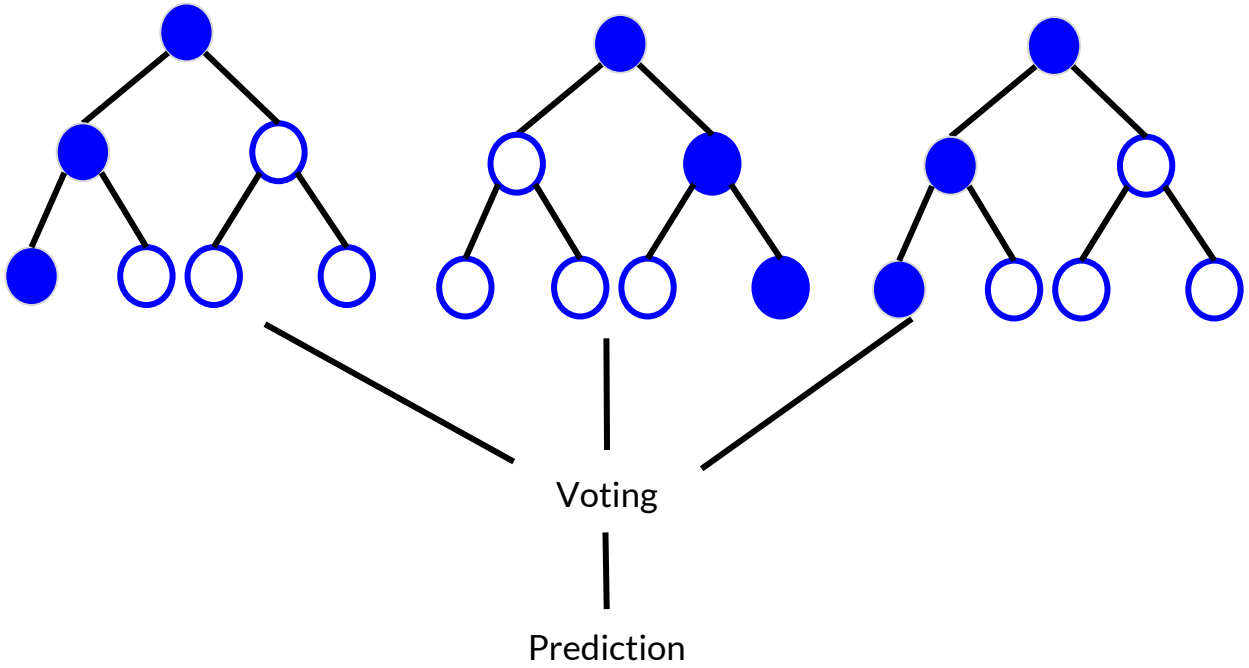
Dropping the nouns

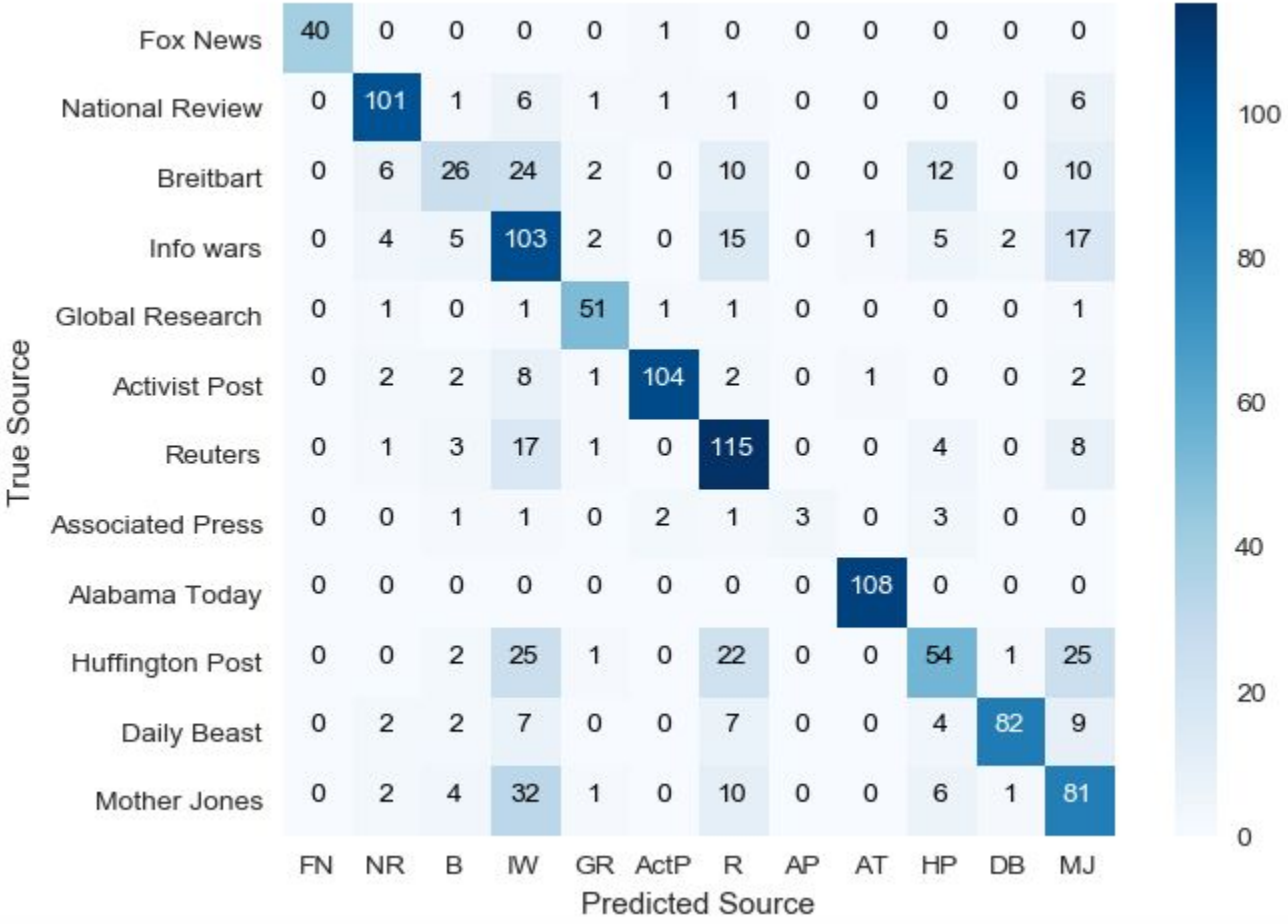
they ~~are a~~ ~~group~~ ~~of~~ 13 ~~highli~~ partisan ~~democrat~~ ~~that~~ make ~~up the mueller team,~~ exclud ~~him,~~ who ~~are~~ tri veri veri hard ~~to~~ ~~frame him,~~ ~~to~~ get ~~him in~~ ~~trouble,~~ ~~when he~~ ~~hasn't~~ done ~~anyth~~ wrong

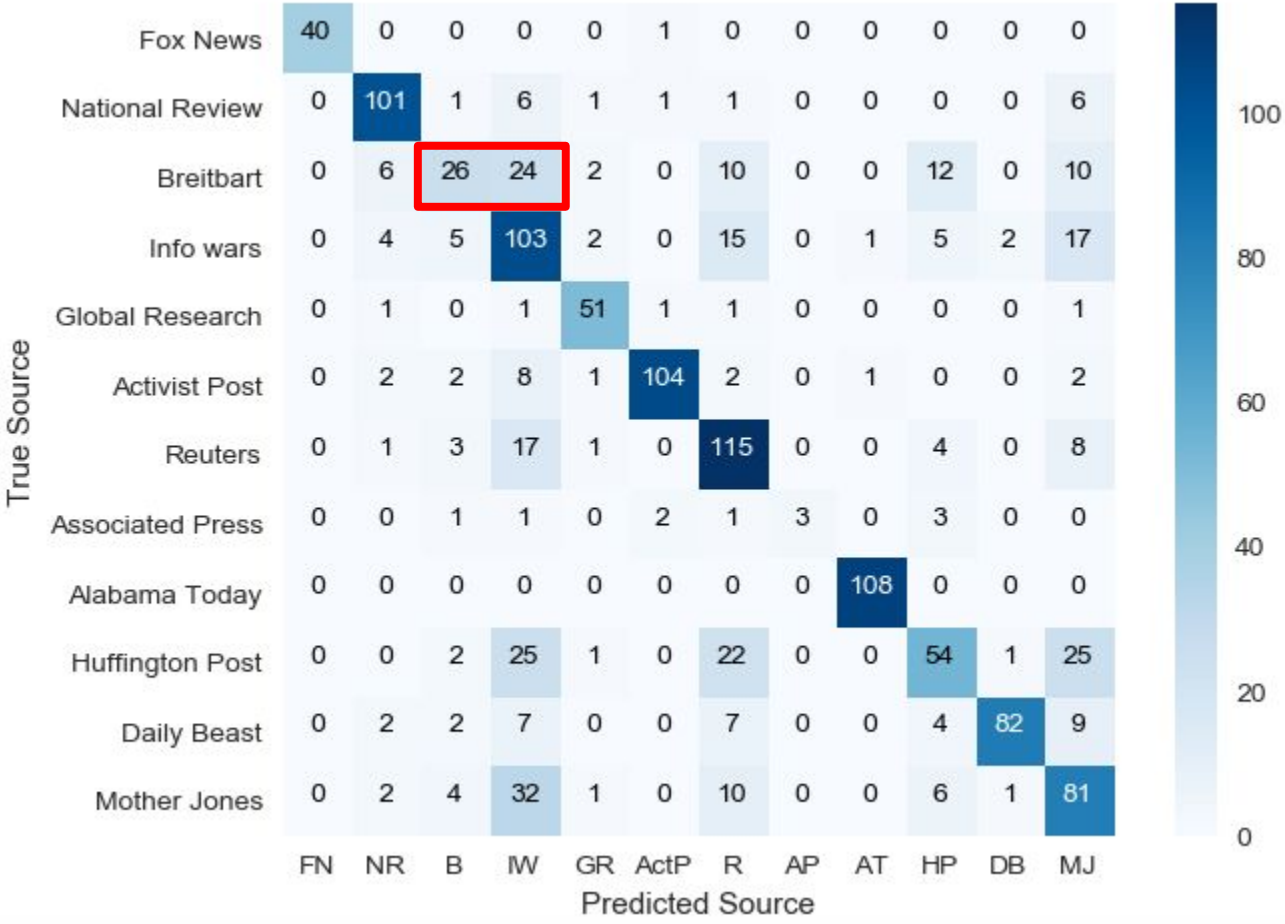
Dropping the nouns

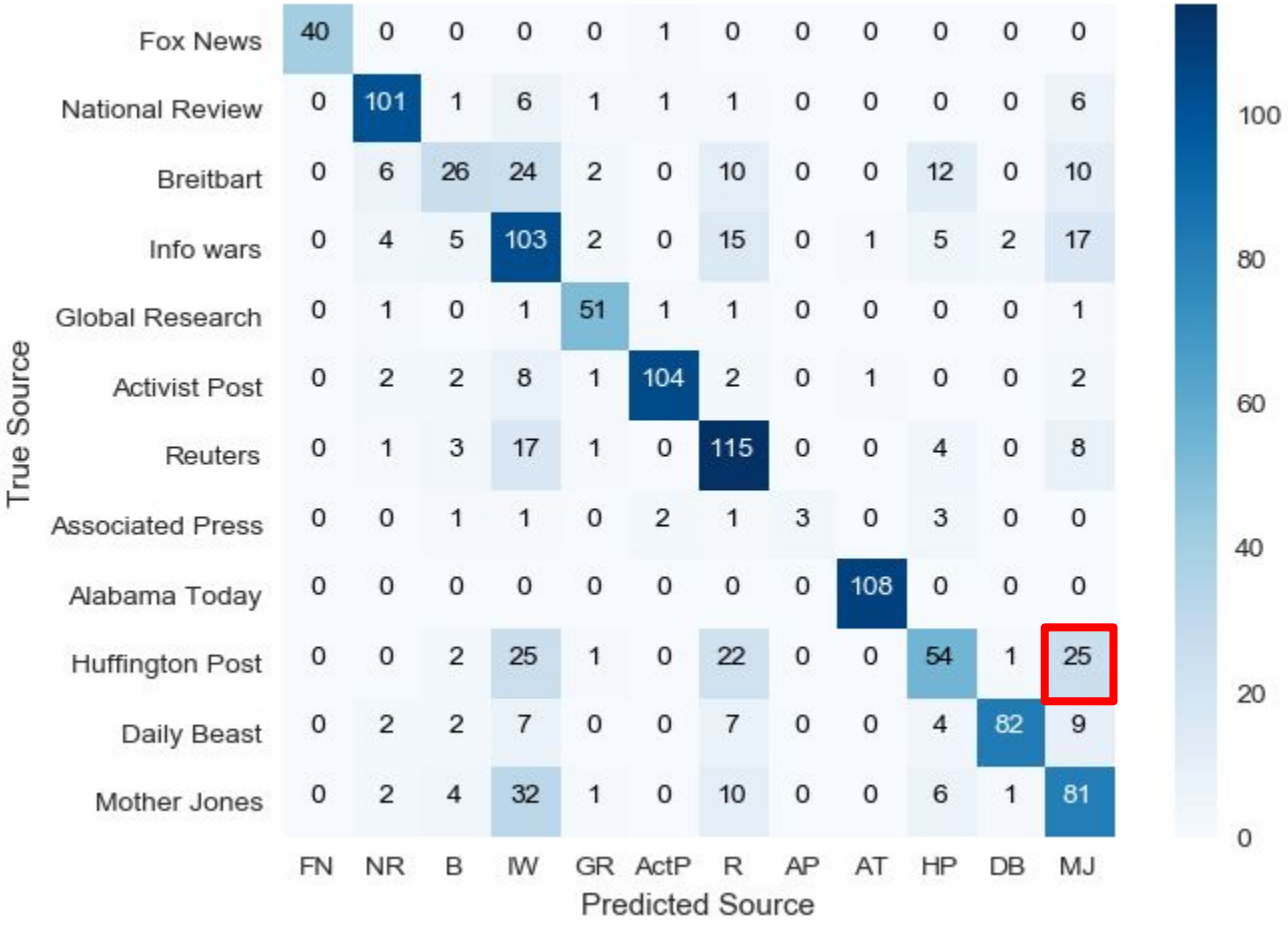
they 13 partisan make exclud who tri very veri hard get done wrong

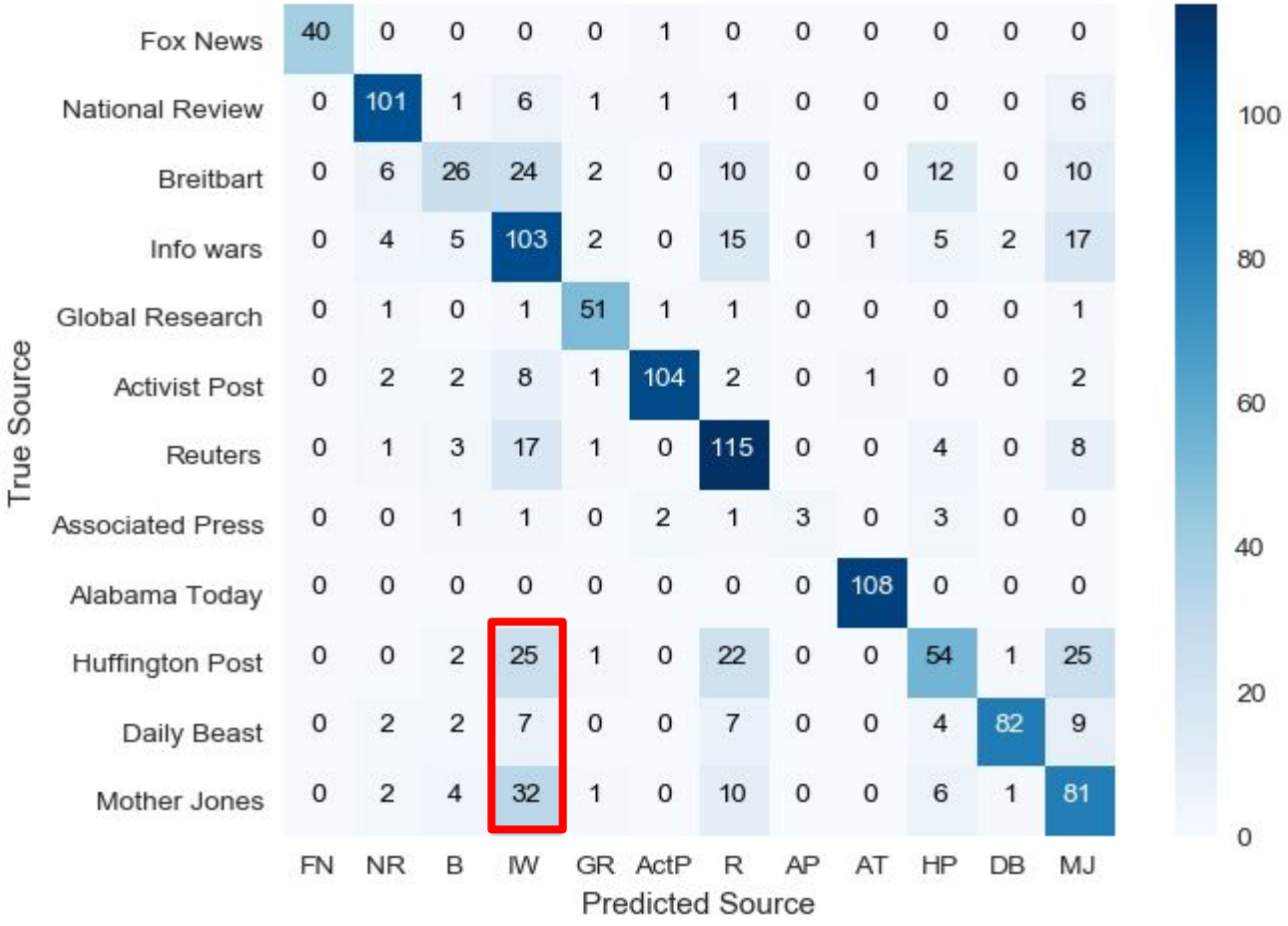
Random Forest

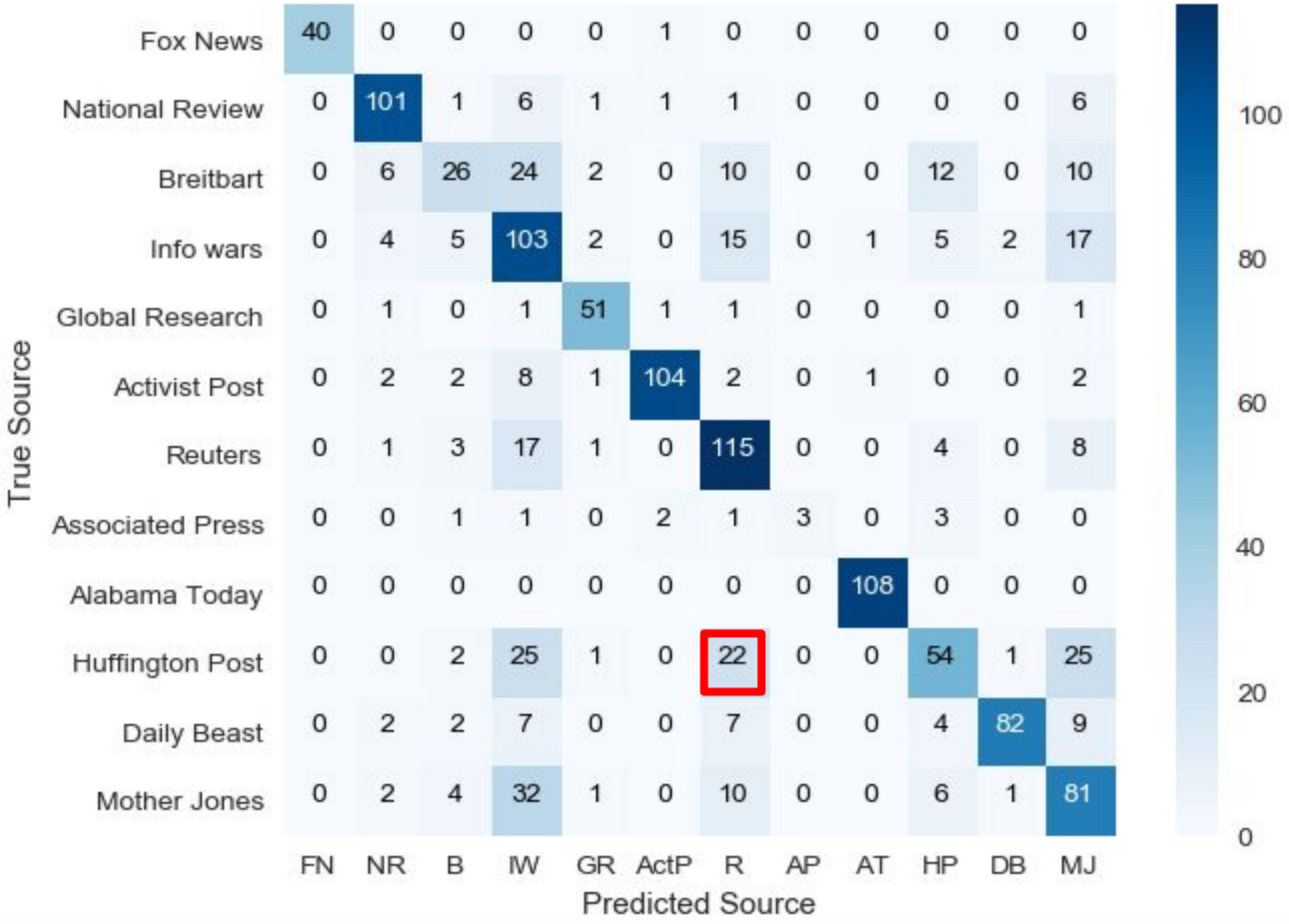


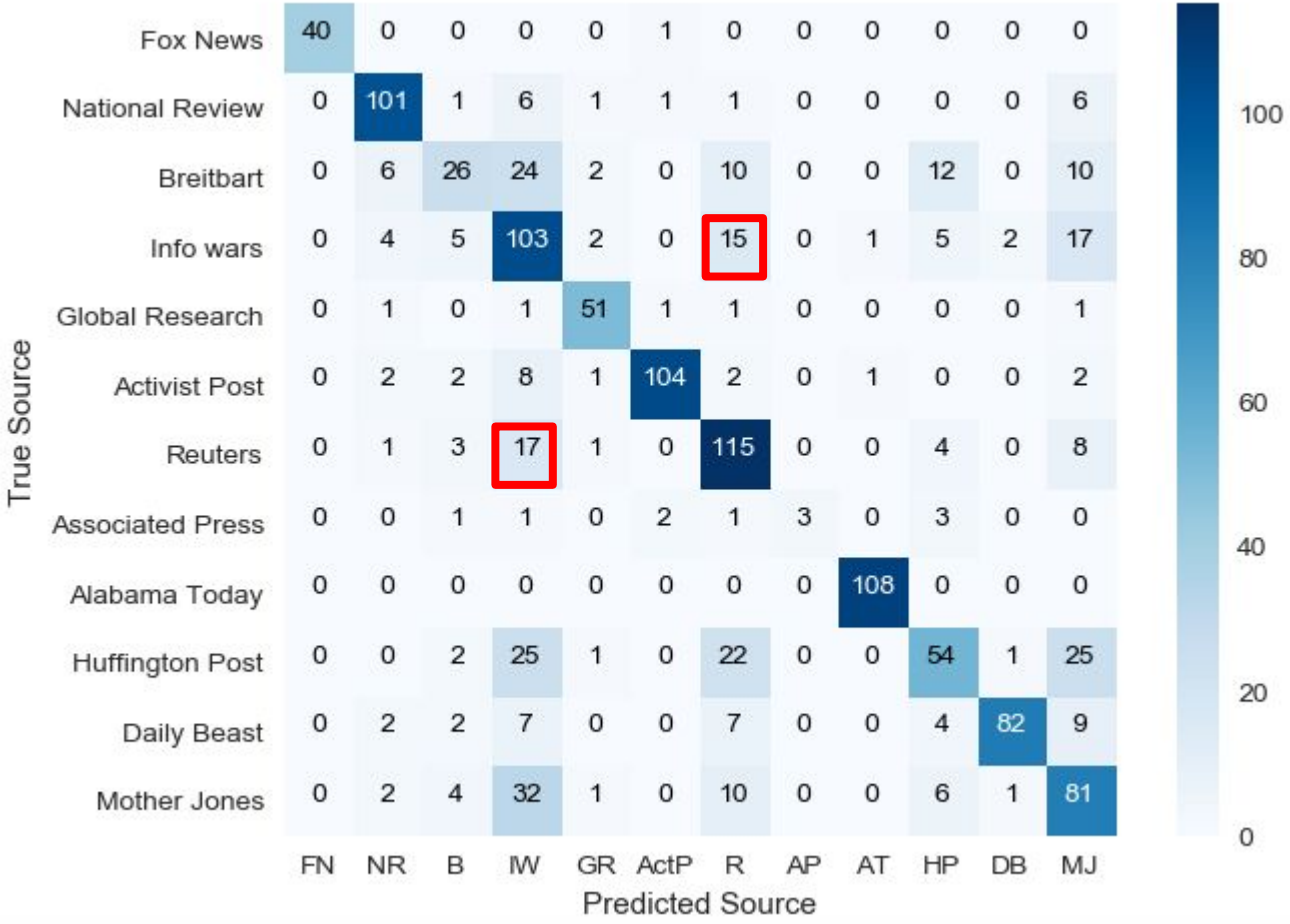

Fox News 40 0 0 0 0 1 0 0 0 0 0 0
National Review 0 101 1 6 1 1 1 0 0 0 0 6
Breitbart 0 6 26 24 2 0 10 0 0 12 0 10
Info wars 0 4 5 103 2 0 15 0 1 5 2 17
Global Research 0 1 0 1 51 1 1 0 0 0 0 1
Activist Post 0 2 2 8 1 104 2 0 1 0 0 2
Reuters 0 1 3 17 1 0 115 0 0 4 0 8
Associated Press 0 0 1 1 0 2 1 3 0 3 0 0
Alabama Today 0 0 0 0 0 0 0 0 108 0 0 0
Huffington Post 0 0 2 25 1 0 22 0 0 54 1 25
Daily Beast 0 2 2 7 0 0 7 0 0 4 82 9
Mother Jones 0 2 4 32 1 0 10 0 0 6 1 81
True Source
FN NR B IW GR ActP R AP AT HP DB MJ
Predicted Source
100
80
60
40
20
0


Fox News 40 0 0 0 0 1 0 0 0 0 0 0
National Review 0 101 1 6 1 1 1 0 0 0 0 6
Breitbart 0 6 26 24 2 0 10 0 0 12 0 10
Info wars 0 4 5 103 2 0 15 0 1 5 2 17
Global Research 0 1 0 1 51 1 1 0 0 0 0 1
Activist Post 0 2 2 8 1 104 2 0 1 0 0 2
Reuters 0 1 3 17 1 0 115 0 0 4 0 8
Associated Press 0 0 1 1 0 2 1 3 0 3 0 0
Alabama Today 0 0 0 0 0 0 0 0 108 0 0 0
Huffington Post 0 0 2 25 1 0 22 0 0 54 1 25
Daily Beast 0 2 2 7 0 0 7 0 0 4 82 9
Mother Jones 0 2 4 32 1 0 10 0 0 6 1 81
True Source
FN NR B IW GR ActP R AP AT HP DB MJ
Predicted Source
100
80
60
40
20
0


Fox News 40 0 0 0 0 1 0 0 0 0 0 0
National Review 0 101 1 6 1 1 1 0 0 0 0 6
Breitbart 0 6 26 24 2 0 10 0 0 12 0 10
Info wars 0 4 5 103 2 0 15 0 1 5 2 17
Global Research 0 1 0 1 51 1 1 0 0 0 0 1
Activist Post 0 2 2 8 1 104 2 0 1 0 0 2
Reuters 0 1 3 17 1 0 115 0 0 4 0 8
Associated Press 0 0 1 1 0 2 1 3 0 3 0 0
Alabama Today 0 0 0 0 0 0 0 0 108 0 0 0
Huffington Post 0 0 2 25 1 0 22 0 0 54 1 25
Daily Beast 0 2 2 7 0 0 7 0 0 4 82 9
Mother Jones 0 2 4 32 1 0 10 0 0 6 1 81
True Source
FN NR B IW GR ActP R AP AT HP DB MJ
Predicted Source
100
80
60
40
20
0


Fox News 40 0 0 0 0 1 0 0 0 0 0 0
National Review 0 101 1 6 1 1 1 0 0 0 0 6
Breitbart 0 6 26 24 2 0 10 0 0 12 0 10
Info wars 0 4 5 103 2 0 15 0 1 5 2 17
Global Research 0 1 0 1 51 1 1 0 0 0 0 1
Activist Post 0 2 2 8 1 104 2 0 1 0 0 2
Reuters 0 1 3 17 1 0 115 0 0 4 0 8
Associated Press 0 0 1 1 0 2 1 3 0 3 0 0
Alabama Today 0 0 0 0 0 0 0 0 108 0 0 0
Huffington Post 0 0 2 25 1 0 22 0 0 54 1 25
Daily Beast 0 2 2 7 0 0 7 0 0 4 82 9
Mother Jones 0 2 4 32 1 0 10 0 0 6 1 81
True Source
FN NR B IW GR ActP R AP AT HP DB MJ
Predicted Source
100
80
60
40
20
0


Fox News 40 0 0 0 0 1 0 0 0 0 0 0
National Review 0 101 1 6 1 1 1 0 0 0 0 6
Breitbart 0 6 26 24 2 0 10 0 0 12 0 10
Info wars 0 4 5 103 2 0 15 0 1 5 2 17
Global Research 0 1 0 1 51 1 1 0 0 0 0 1
Activist Post 0 2 2 8 1 104 2 0 1 0 0 2
Reuters 0 1 3 17 1 0 115 0 0 4 0 8
Associated Press 0 0 1 1 0 2 1 3 0 3 0 0
Alabama Today 0 0 0 0 0 0 0 0 108 0 0 0
Huffington Post 0 0 2 25 1 0 22 0 0 54 1 25
Daily Beast 0 2 2 7 0 0 7 0 0 4 82 9
Mother Jones 0 2 4 32 1 0 10 0 0 6 1 81
True Source
FN NR B IW GR ActP R AP AT HP DB MJ
Predicted Source
100
80
60
40
20
0


Fox News 40 0 0 0 0 1 0 0 0 0 0 0
National Review 0 101 1 6 1 1 1 0 0 0 0 6
Breitbart 0 6 26 24 2 0 10 0 0 12 0 10
Info wars 0 4 5 103 2 0 15 0 1 5 2 17
Global Research 0 1 0 1 51 1 1 0 0 0 0 1
Activist Post 0 2 2 8 1 104 2 0 1 0 0 2
Reuters 0 1 3 17 1 0 115 0 0 4 0 8
Associated Press 0 0 1 1 0 2 1 3 0 3 0 0
Alabama Today 0 0 0 0 0 0 0 0 108 0 0 0
Huffington Post 0 0 2 25 1 0 22 0 0 54 1 25
Daily Beast 0 2 2 7 0 0 7 0 0 4 82 9
Mother Jones 0 2 4 32 1 0 10 0 0 6 1 81
True Source
FN NR B IW GR ActP R AP AT HP DB MJ
Predicted Source
100
80
60
40
20
0

Applications

- Pipeline to scrape several sources at once.

- Automatically differentiate writing styles.

Email: nickthomas3245@gmail.com

Linkedin: nicholas-thomas-zingo3245

Github: Zingo3245