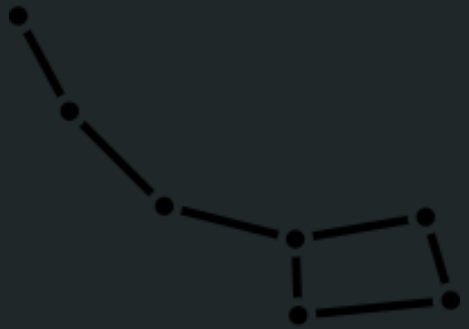
Team: Polaris



Political stance classification of news-articles

By: Aahana, Allen, Byeongsu, Israel, Jacob, and Tessa
TAs: Zahra and Shaonan
Mentors: Milan and Vahid

neuromatch academy

July 2022
Pod: ajiaco
Megapod: nunki

Background & Aim

- Political bias in media plays a detrimental role in shaping public opinion.
- Media bias ratings are mostly based on subjective opinion of individuals.
- However, a preference blind and context dependent, approach is better suited for this classification.

We ask the following questions in this project:

1. Can we use machine learning to infer political stance of news articles?
2. Are specific words related to any political stance?

“BIG NEWS” Dataset

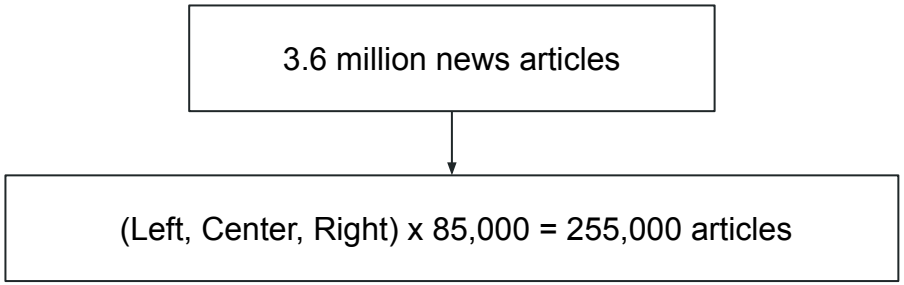
Consists of 3,689,229 English news articles on politics, gathered from 11 United States (US) media outlets covering a broad ideological spectrum.

	Daily Kos	HPO	CNN	WaPo	NYT	USA Today	AP	The Hill	TWT	FOX	Breitbart
Ideology	L	L	L	L	L	C	C	C	R	R	R
# articles	100,828	241,417	64,988	198,529	173,737	170,737	279,312	322,145	243,181	330,166	206,512
# words	738.7	729.9	655.7	803.2	599.4	691.7	572.3	426.3	522.7	773.5	483.5

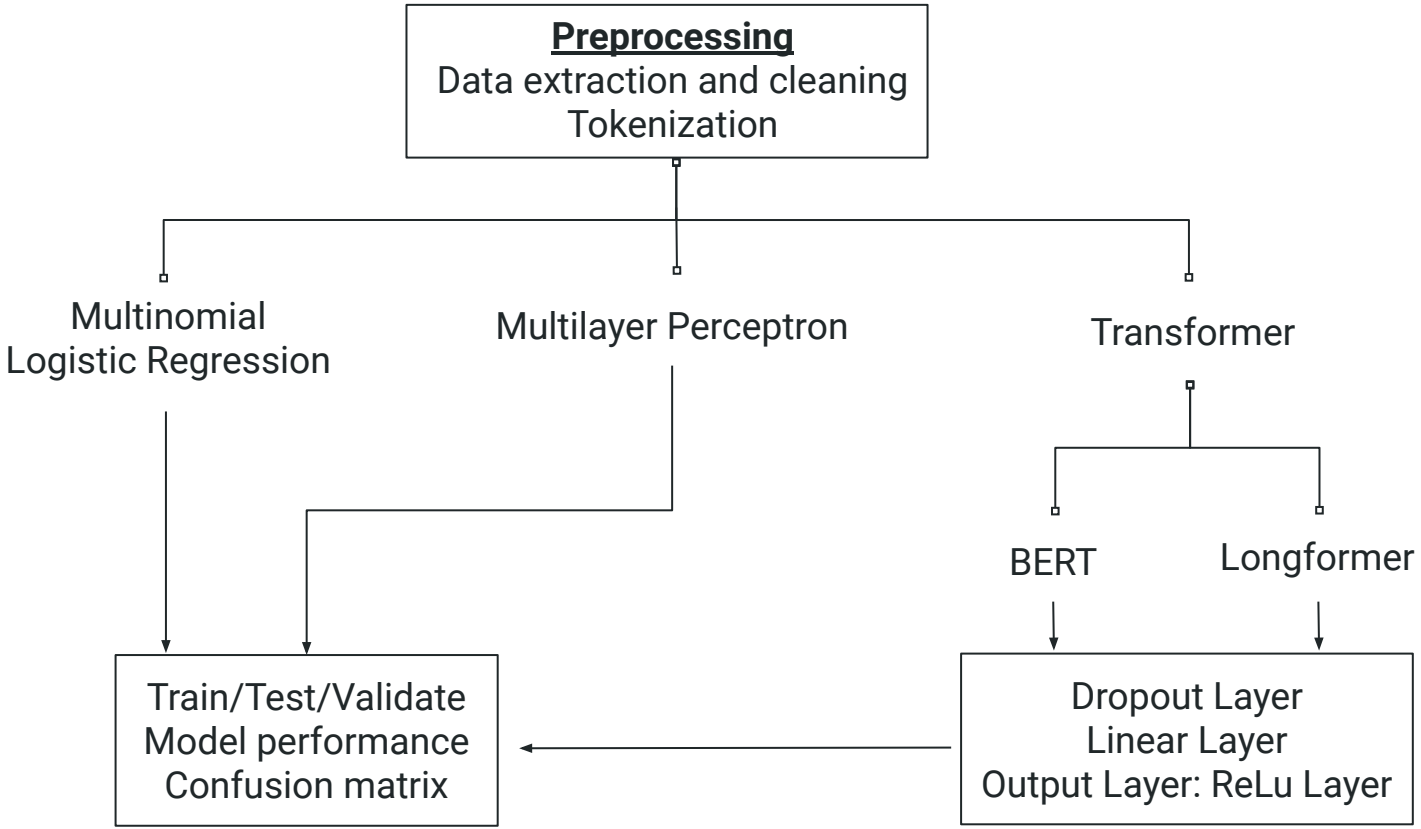
Liu et.al 2022, NAACL

Issues encountered:

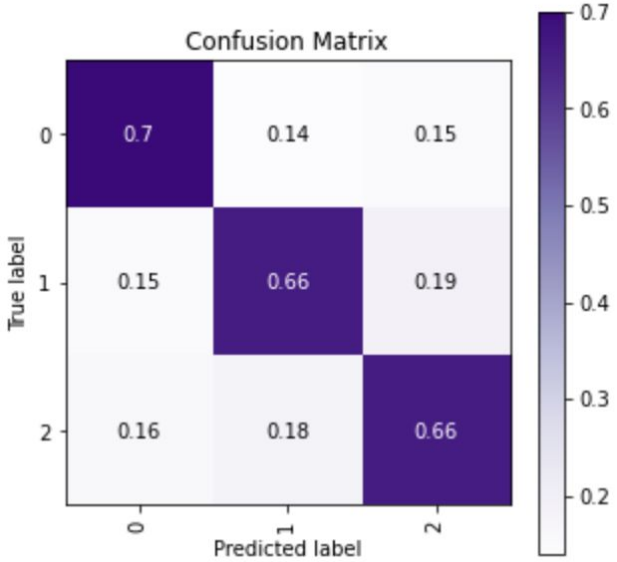
- Storage limitations
- GPU limitations
- Parallel processing with Dask
- Chunking into smaller datasets



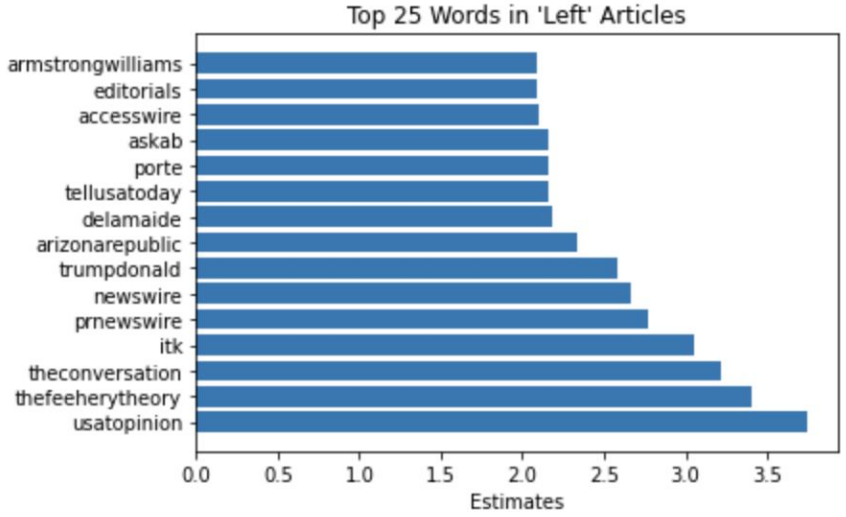
Methods

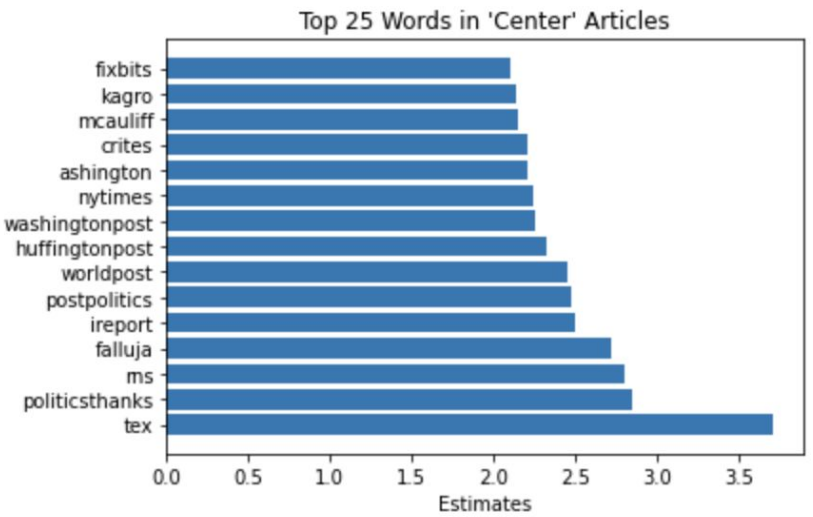

Preprocessing
Data extraction and cleaning
Tokenization
Multinomial
Logistic Regression
Multilayer Perceptron
Transformer
BERT
Longformer
Train/Test/Validate
Model performance
Confusion matrix
Dropout Layer
Linear Layer
Output Layer: ReLu Layer

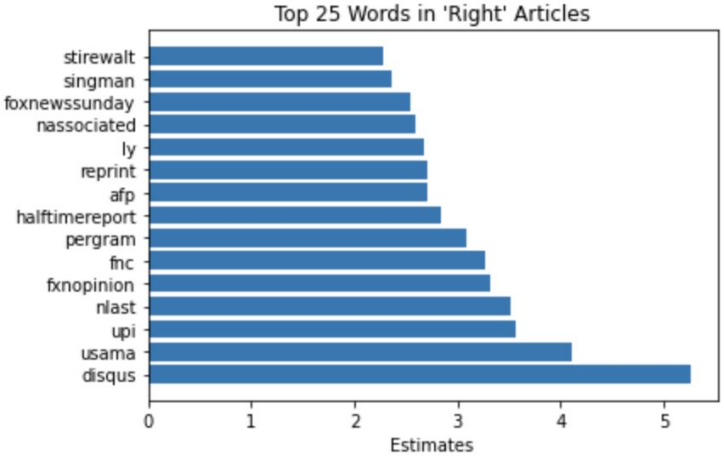
Results: Logistic Regression



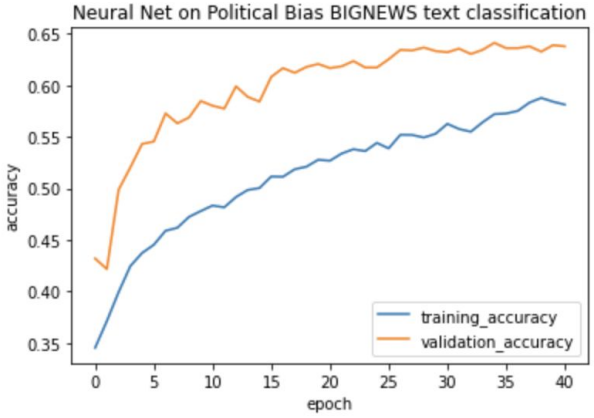
	precision	recall	f1-score	support
0	0.70	0.69	0.70	35100
1	0.66	0.68	0.67	35100
2	0.66	0.65	0.66	35100
accuracy			0.67	105300
macro avg	0.67	0.67	0.67	105300
weighted avg	0.67	0.67	0.67	105300

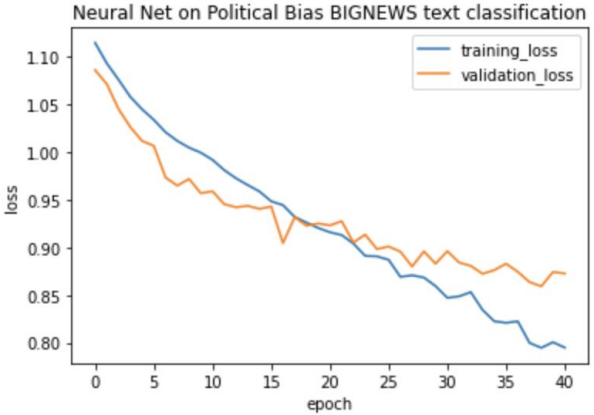

Top 25 Words in 'Left' Articles
armstrongwilliams
editorials
accesswire
askab
porte
tellusatoday
delamaide
arizonarepublic
trumpdonald
newswire
prnewswire
itk
theconversation
thefeeherytheory
usatopinion
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Estimates


Top 25 Words in 'Center' Articles
fixbits
kagro
mcauliff
crites
ashington
nytimes
washingtonpost
huffingtonpost
worldpost
postpolitics
ireport
falluja
ms
politicsthanks
tex
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Estimates


Top 25 Words in 'Right' Articles
stirewalt
singman
foxnewssunday
nassociated
ly
reprint
afp
halftimereport
pergram
fnc
fxnopinion
nlast
upi
usama
disqus
0
1
2
3
4
5
Estimates

Context-Insensitive MLP Using Pretrained FastText Embeddings





Test Accuracy: 64%

- Hidden Layers = 3
- Hidden Size = 2000
- Dropout = 60%
- L2 Regularization = 0.000005
- Learning Rate = 0.0001
- Finetuning FastText pretrained embeddings
- Slightly worse than logistic regression

Layers of the neural network model

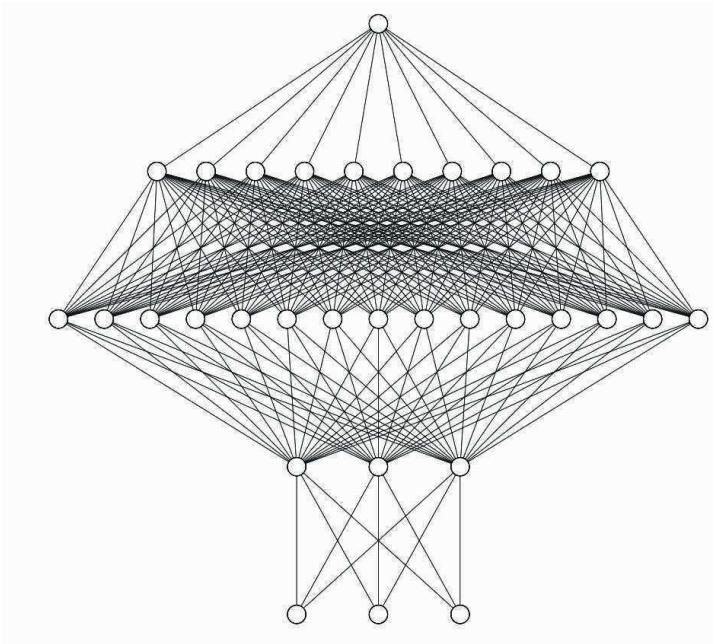
Tokenization
(512 or 1000)

BERT or Longformer
(512 or 1000 → 768)

Linear (Pooling) Layer
(768 → 3)

ReLU Layer
(3 → 3)

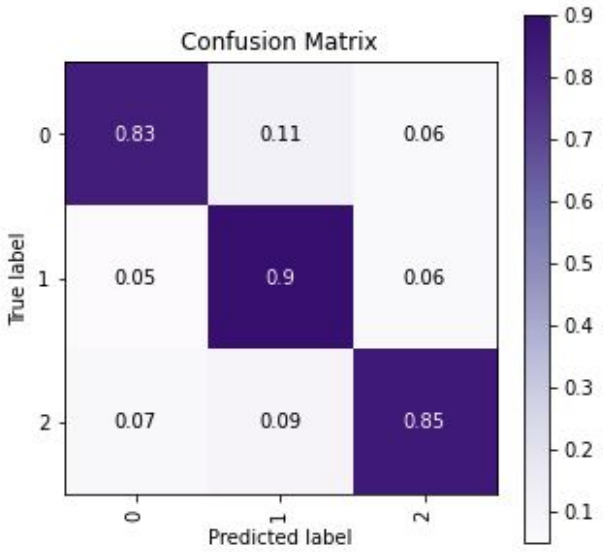
News article



Left Center Right

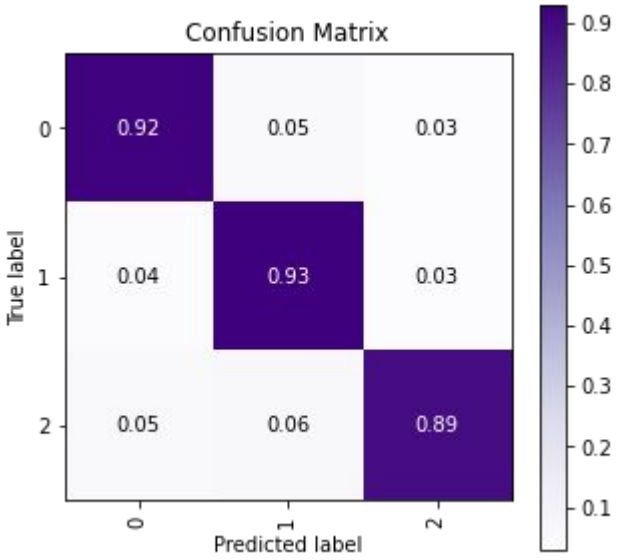
Results: Transformer

BERT



Epochs = 5

Longformer



Epochs = 2

Conclusion/Discussion

1. Can we use machine learning to infer political stance of news articles? **Yes**
2. Are specific words related to any political stance? **Maybe**

Key points:

- Logistic regression was able to predict news article leaning with an accuracy > 70%
- Simple context insensitive NN using pre-trained FastText embeddings only had 55-62% accuracy
- Longformer model was the more effective than BERT in classifying political leaning

Limitations & Future Directions

- Classification only based on the first few sentences of articles
  → Sliding window approach over long documents
- News outlet ≠ “Political stance”
  → Alternative dataset?
  → Feature extraction to characterize features most predictive of different political leanings