Can We Detect Political Bias in Tweets?

Capstone NLP Project by Sarah Zoeller

Business Case

Given the recent political context, various organizations are exploring the use of machine learning to fight political bias in text content such as news articles and social media posts. This project aims to explore if a machine learning or deep learning algorithm can distinguish between content from left leaning politicians vs right leaning politicians.

So...can we detect political bias in tweets?

The Data

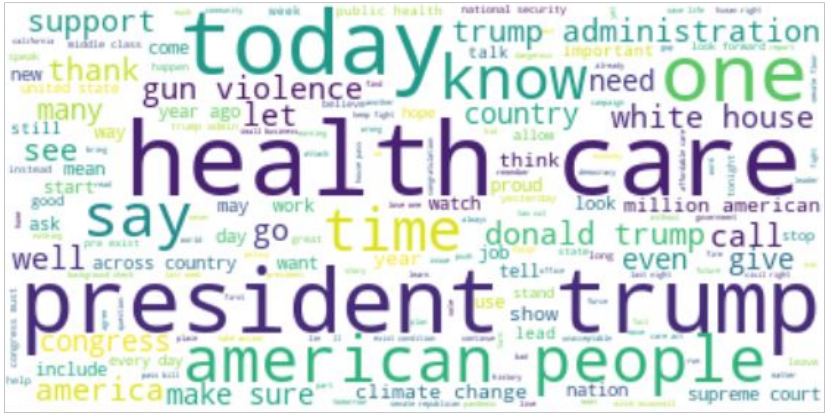
- Data was sourced from Twitter using Twint
- Included over 400,000 tweets from 500+ congressional members and politicians from January 2016 - February 2021
- Data was labeled as left or right biased **based on the political party of the tweeter - *See Limitations***
- 75% of data was class “D” (Democrat), 25% class “R” (Republican)
- Train test split 70/30 train/test

Exploratory Data Analysis – Word Clouds

Right Wing Tweets - Word Cloud

- Words that are less topic/policy oriented, temporal - today, time, need

Left Wing Tweets - Word Cloud



- Topic oriented: health care, gun violence
- Both mention other side frequently: ie President Trump, democrat

Exploratory Data Analysis - Statistics

Frequent tweeters from the Right:

```
Rubio, Marco
Cruz, Ted
Cornyn, John
Paul, Rand
Blackburn, Marsha

Accounting for a total of 28027 tweets
```

Frequent tweeters from the Left:

```
Padilla, Alex
Murphy, Chris
Schumer, Chuck
Durbin, Richard
Klobuchar, Amy

Accounting for a total of 42099 tweets
```

Popular tweeters from the Right:

```
Jordan, Jim
Pompeo, Mike
Pence, Mike
Nunes, Devin
McCarthy, Kevin

Accounting for an average of 10524.96 likes per tweet
```

Popular tweeters from the Left:

```
Obama, Barack
Biden, Joe
Schiff, Adam
Lewis, John
Clinton, Hillary

Accounting for an average of 46698.36 likes per tweet
```

Exploratory Data Analysis – LDA

Right Topics - National Security/International Relations, Economy, Elections

```
[(0,
  '0.015*"today" + 0.013*"president" + 0.013*"year" + 0.013*"day" + 0.012*"great" + 0.012*"thank" + 0.012*"america" + 0.011*"li
fe" + 0.011*"nation" + 0.011*"honor"'),
 (1,
  '0.013*"border" + 0.011*"china" + 0.011*"act" + 0.010*"security" + 0.009*"must" + 0.008*"support" + 0.007*"bill" + 0.007*"la
w" + 0.007*"protect" + 0.007*"right"'),
 (2,
  '0.018*"great" + 0.017*"today" + 0.016*"join" + 0.011*"discuss" + 0.010*"thank" + 0.009*"morning" + 0.009*"texas" + 0.009*"wa
tch" + 0.009*"tonight" + 0.009*"I"'),
 (3,
  '0.015*"vote" + 0.014*"democrat" + 0.012*"president" + 0.011*"house" + 0.009*"election" + 0.009*"people" + 0.008*"impeachmen
t" + 0.008*"say" + 0.008*"senate" + 0.007*"trump"'),
 (4,
  '0.015*"job" + 0.014*"work" + 0.012*"need" + 0.009*"business" + 0.009*"get" + 0.008*"help" + 0.008*"people" + 0.008*"economy"
+ 0.007*"america" + 0.007*"year"')]
```

Left Topics - Healthcare, Work and Taxes, Elections

```
[(0,
  '0.016*"pay" + 0.011*"worker" + 0.011*"tax" + 0.011*"work" + 0.010*"make" + 0.009*"job" + 0.009*"climate" + 0.009*"family" +
0.008*"need" + 0.008*"cut"'),
 (1,
  '0.016*"vote" + 0.014*"I" + 0.013*"get" + 0.013*"make" + 0.012*"today" + 0.011*"day" + 0.009*"gun" + 0.008*"go" + 0.008*"tim
e" + 0.008*"work"'),
 (2,
  '0.014*"woman" + 0.013*"country" + 0.011*"right" + 0.011*"today" + 0.010*"fight" + 0.010*"year" + 0.009*"nation" + 0.009*"fam
ily" + 0.009*"america" + 0.009*"stand"'),
 (3,
  '0.030*"health" + 0.029*"care" + 0.015*"need" + 0.012*"act" + 0.011*"people" + 0.011*"million" + 0.011*"right" + 0.010*"must"
+ 0.010*"family" + 0.009*"bill"'),
 (4,
  '0.028*"president" + 0.010*"donald" + 0.008*"must" + 0.007*"election" + 0.007*"people" + 0.007*"republican" + 0.007*"say" +
0.006*"congress" + 0.006*"house" + 0.006*"hillary"')]
```

Modeling Process and Results

Machine Learning Models:

- Logistic Regression
- Random Forest
- Naive Bayes

Best Model After Tuning:

	Train Scores LogisticRegression				Test Scores LogisticRegression			
	precision	recall	f1-score	support	precision	recall	f1-score	support
D	0.87	0.79	0.83	212231.00	0.87	0.79	0.83	91216.00
R	0.50	0.65	0.56	69181.00	0.49	0.64	0.55	29390.00
accuracy	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
macro avg	0.69	0.72	0.70	281412.00	0.68	0.71	0.69	120606.00
weighted avg	0.78	0.75	0.76	281412.00	0.78	0.75	0.76	120606.00

Deep Learning Models:

- Recurrent Neural Network
  - GRU
  - LSTM

Best Model After Tuning:

- **80% accuracy, 0.75 F1 Score**

	Train Scores Neural Network				Test Scores Neural Network			
	precision	recall	f1-score	support	precision	recall	f1-score	support
0	0.95	0.82	0.88	212231.00	0.92	0.80	0.86	91216.0
1	0.61	0.86	0.72	69181.00	0.56	0.79	0.65	29390.0
accuracy	0.83	0.83	0.83	0.83	0.80	0.80	0.80	0.8
macro avg	0.78	0.84	0.80	281412.00	0.74	0.79	0.75	120606.0
weighted avg	0.87	0.83	0.84	281412.00	0.83	0.80	0.81	120606.0

Prediction Function

```
predict_class_NN()
```

```
Please enter a tweet:
Joe Biden’s been President for 43 days. It’s only going to get worse.

'Potentially Right Biased'
```

```
predict_class_NN()
```

```
Please enter a tweet:
Conservative Dems have fought so the Biden admin sends fewer & less generous relief checks than the Trump admin did.  It’s a mo
ve that makes little-to-no political or economic sense, and targets an element of relief that is most tangibly felt by everyday
people. An own-goal.

'Potentially Left Biased'
```

Post Modeling EDA

Correctly classified ”R”:

['It’s a charade anyway – Democratic leader Chuck Schumer controls Nelson’s vote. Senator Nelson’s entire campaign is funded by the Democrat leaders and special interests in Washington. Schumer and Pelosi own Nelson’s vote… he doesn’t even have the option of voting with Florida.']

Correctly classified “D”:

['Don’t let the NRA fool you: an overwhelming 97% of Americans support universal background checks to confront gun violence. We owe it to every victim, survivor, their families, and our communities to be stronger and louder than the gun lobby. https://t.co/XzBXuuZ6Cd']

Incorrectly classified as “R”:

['Even conservatives acknowledge that Turkey is becoming an autocracy. Erdogan is setting the country back 100 years. https://t.co/OJyUAeeBuy']

Incorrectly classified as “D”:

['Trump: Responsibly Restrained on Foreign Policy. https://t.co/hjZGyTDWwa']

Conclusions and Recommendations

- While there appears to be a detectable difference in the way right or left leaning politicians tweet, the model does not perfectly distinguish between the two and could use further tuning.
- Model could be more accurate with added classes, such as “neutral”, “mild right bias”, “mild left bias”.
- If the model were to be used, it could label articles/posts as “May Contain Political Bias” for the content that is determined to be far right or left (based on their probabilities).
- This would allow readers to at least be aware that there may be bias.

Limitations and Next Steps

Limitations

- Labeling: Data labeling was based on political party of the tweeter. Party of tweeter does not necessarily serve as a proxy for political bias
- Class distributions: Some politicians who tweet more frequently were more represented in the data than others. This could introduce bias into the dataset if said politicians use a certain rhetoric
- Inclusion/exclusion of political figures: Dataset was restricted to those figures on twitter. Politicians who are not on the platform were not included.

Next Steps

- Explore including prominent figures into dataset who are not on Twitter
- Set minimum and maximum threshold for tweets to be included in dataset so there is not over/under representation of certain tweeters
- Removal/inclusion of tweets based on their content
- Implementation of document embeddings
- Expansion of classes (neutral, far right, far left, mild right, mild left).

Sources

- TriageCancer.org
- Twitter
- Twint Project
- Stanford NLP Group - GloVe

Thank You

Github: https://github.com/swzoeller