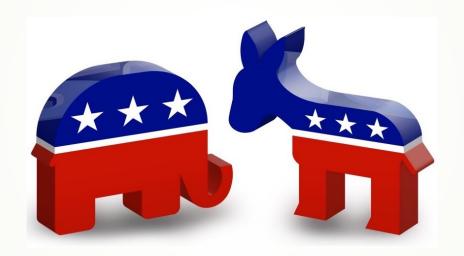
# Differentiating Between Political Party Subreddits



Ben Moss

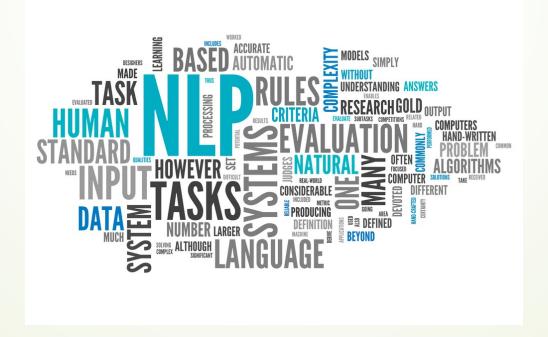
Data Science Immersive Program

General Assembly

January, 2024

#### Goals

- Create Natural Language Processing Models the predict subreddit content origin
- Choose the best ones suited for our goal which in this case will be *sensitivity*: The probability our model will correctly predict a title entry to be in the 'democrats' subreddit (positive class, denoted '1')
- The lines are more blurred than one may think!



#### **Data Gathering**

Worked with PRAW retrieving posts over several days

Loaded data and performed EDA



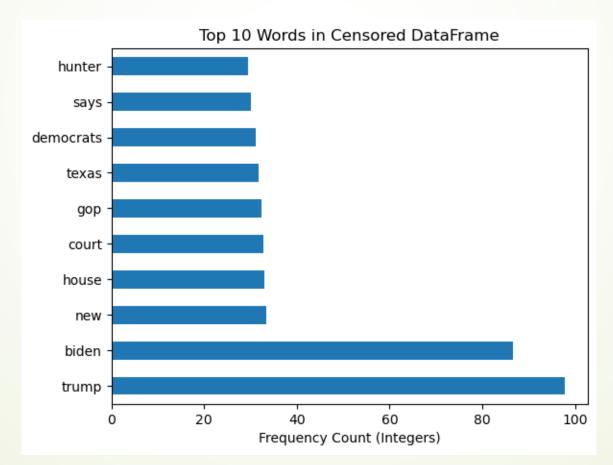
subreddit

Dem: 1 0.528996

Rep: 0 0.471004

# **Modeling Pipeline**

- First and foremost, Regex and censor functions utilized to clean data, literally!
- Several models were compared for efficacy (optimality along with standard classification metrics). The two top performing models were selected



# Deciding Upon NLP Models

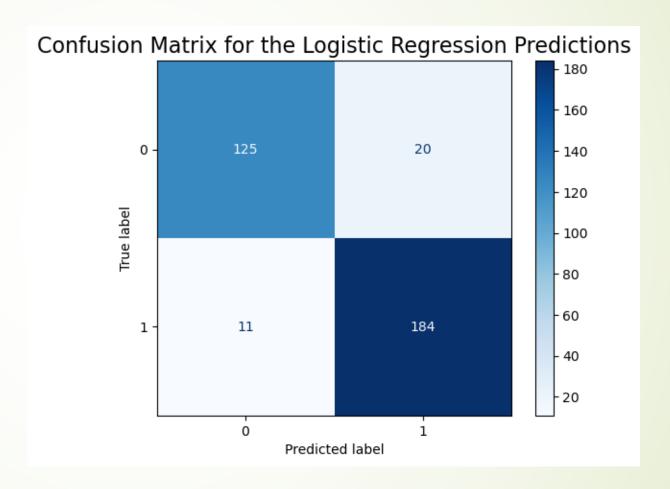
■Best accuracy – optimal fit between test and train

Classification Metrics

| Model          | CVEC Logistic<br>Regression | TVEC SVM     |
|----------------|-----------------------------|--------------|
| Train Accuracy | 0.986                       | <u>0.995</u> |
| Test Accuracy  | 0.909                       | <u>0.921</u> |

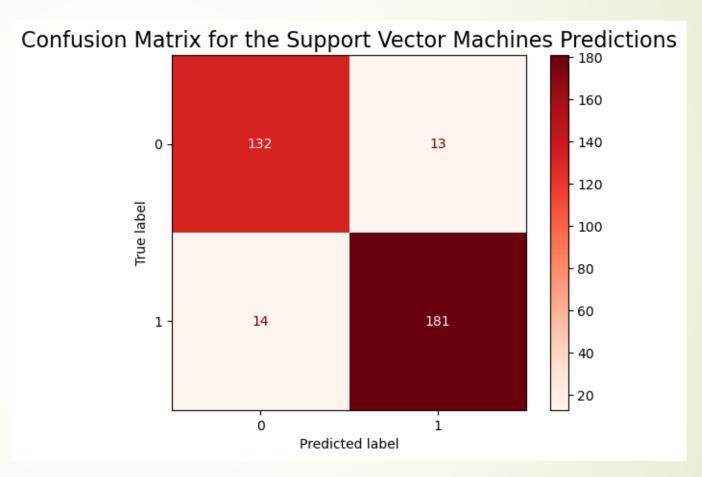
#### **Classification Metrics**

- ► LR (CVEC):
- Sensitivity | 0.944
- ► Specificity | 0.862
- ► Accuracy | 0.909
- Precision | 0.902
- Miscalculation Rate | 0.091



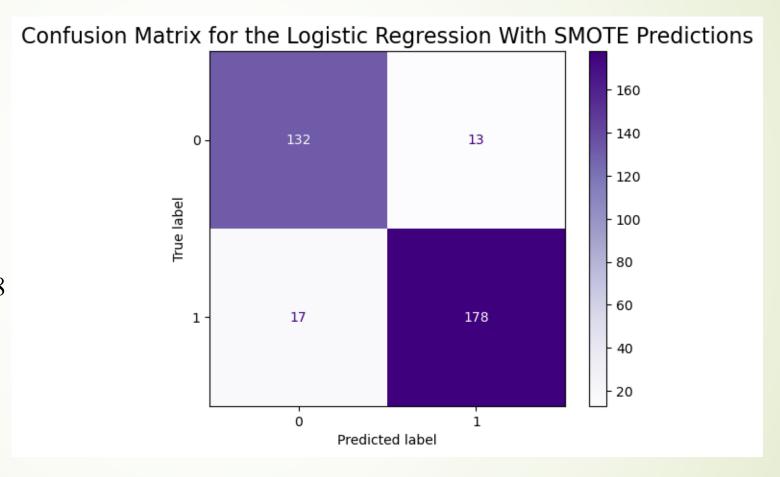
#### Classification Metrics (Continued)

- SVM (TVEC):
- Sensitivity | 0.928
- Specificity | 0.91
- ► Accuracy | 0.921
- Precision | 0.933
- Miscalculation Rate | 0.079



#### **BONUS: Logistic Regression with SMOTE**

- LR (CVEC) SMOTE:
- Sensitivity | 0.913
- Specificity | 0.91
- Accuracy | 0.912
- Precision | 0.932
- ► Miscalculation Rate | 0.088



## Side-by-Side Classification Metric Comparison

| Model               | LR (CVEC)    | SVM (TVEC)   | LR (CVEC) SMOTE |
|---------------------|--------------|--------------|-----------------|
| Sensitivity         | <u>0.944</u> | 0.928        | 0.913           |
| Specificity         | 0.862        | <u>0.91</u>  | <u>0.91</u>     |
| Accuracy            | 0.909        | <u>0.921</u> | 0.912           |
| Precision           | 0.902        | <u>0.933</u> | 0.932           |
| Miscalculation Rate | 0.091        | <u>0.079</u> | 0.088           |

#### Recommendations

- Exclude comments (formatting issue)
- ► Import more of one set to form an even split (when one is more active)
- Use not only optimal model but ones that return the desired classification metrics for intended purposes (sensitivity in this case)
- In this case, avoid random sampling, utilize selective sampling (best subset)



# Conclusion

- Optimal Model: SVM (TVEC)
- Goal of our model: prioritize sensitivity, this result was achieved
- In this study, SVM outperformed logistic regression.
- Successfully able to differentiate between political supporters of two distinct parties.

### **Sources**

Course work, lessons, Reddit, PRAW, .sklearn documentation centers