Differentiating Political Party Subreddits

Ben Moss

Data Science Immersive Program

General Assembly

January, 2024

Goals

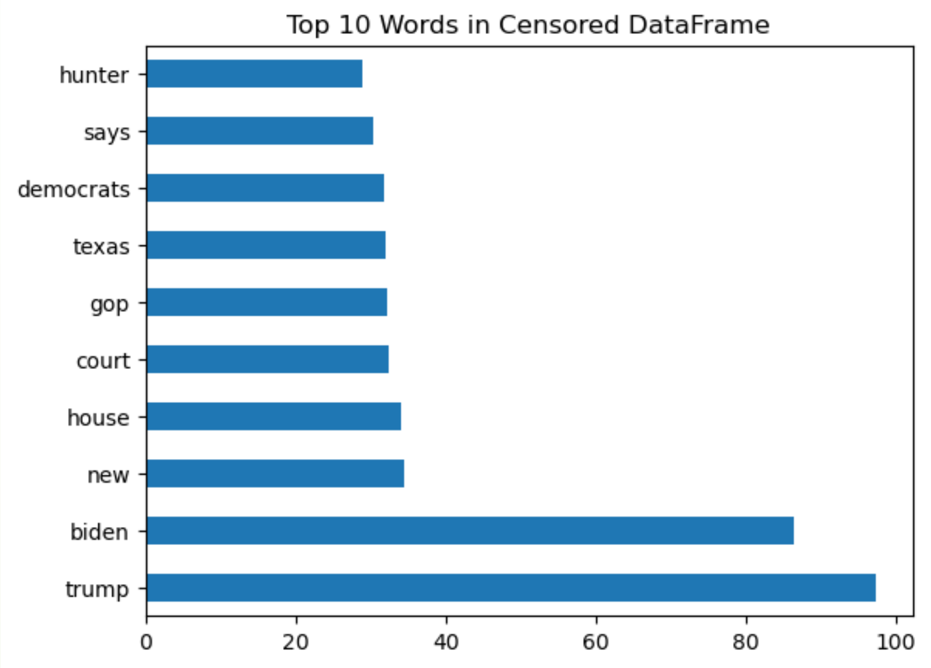
- Create Natural Language Processing Models the predict subreddit content origin
- Choose the best ones suited for our goal which in this case will be *sensitivity :* The probability my model will correctly predict a title entry to be in the \`democrats\` subreddit (positive class, denoted \`1\`)
- ***The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

Data Gathering

- Worked with PRAW retrieving posts over several days
- Loaded data and performed EDA

Modeling Pipeline

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



Deciding Upon NLP Models

Best Accuracy and optimal agreement between test and train

CVEC with logistic regression

Train Accuracy Score | 0.986

Test Accuracy Score | 0.909

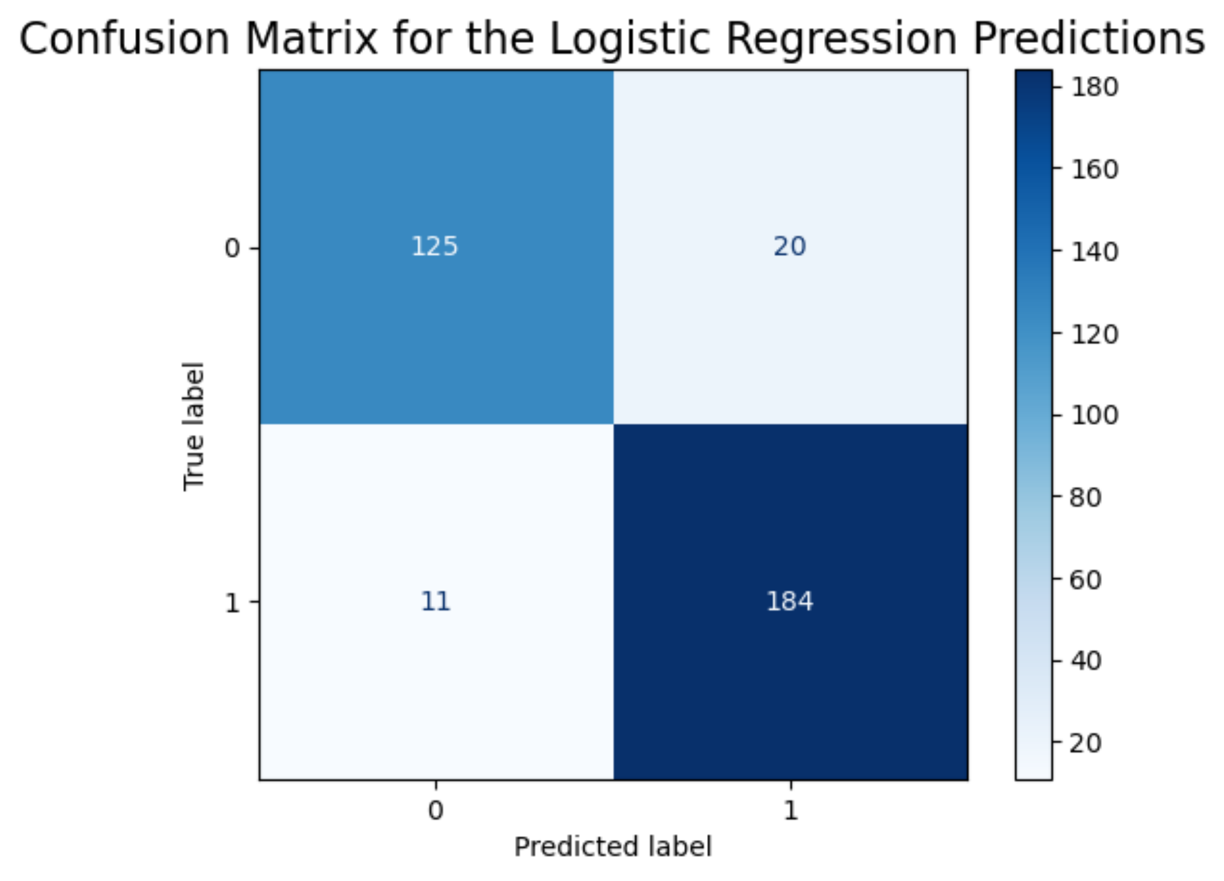
TVEC SVM

Train Accuracy Score | 0.995

Test Accuracy Score | 0.921

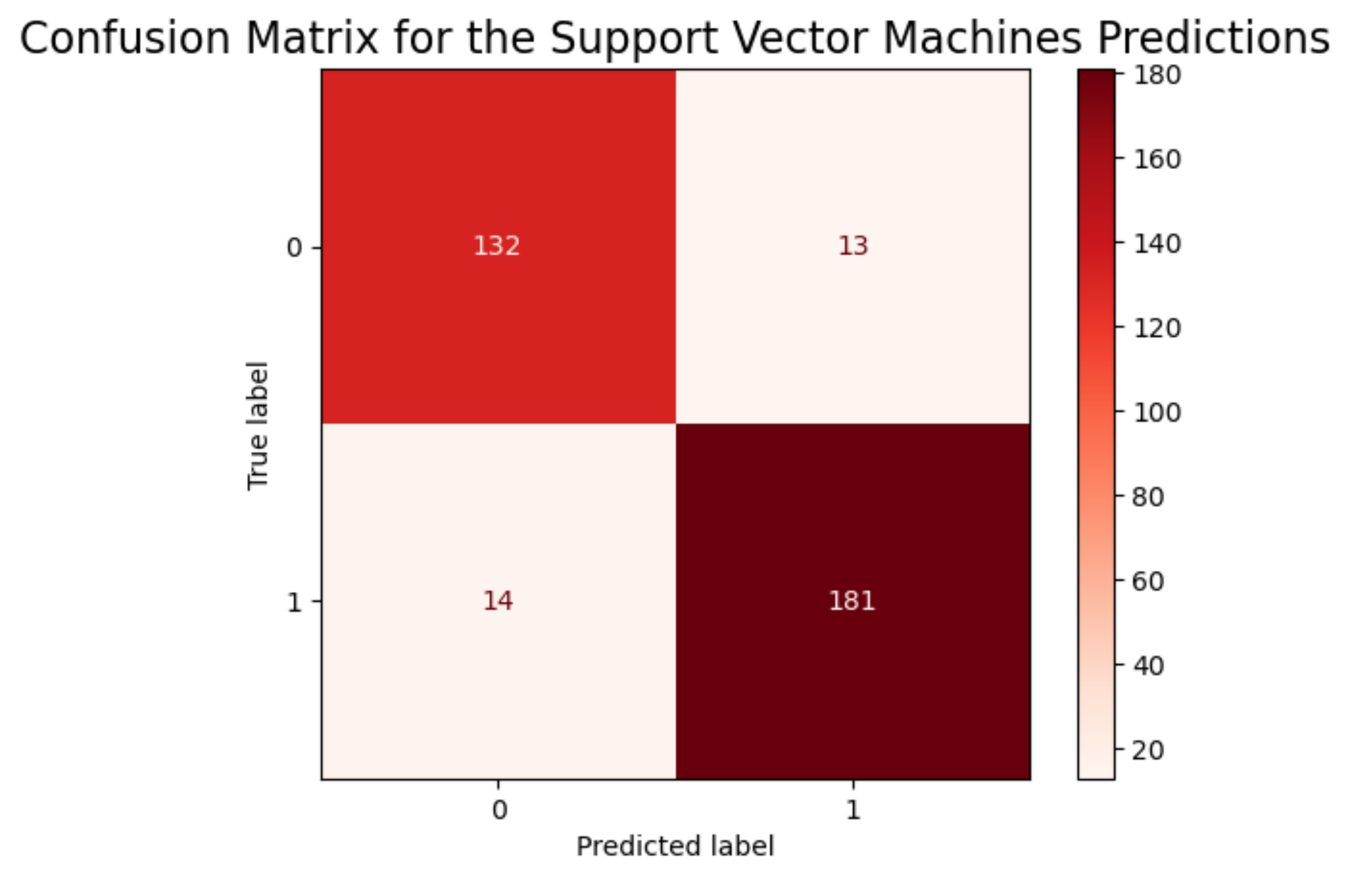
Classification Metrics

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



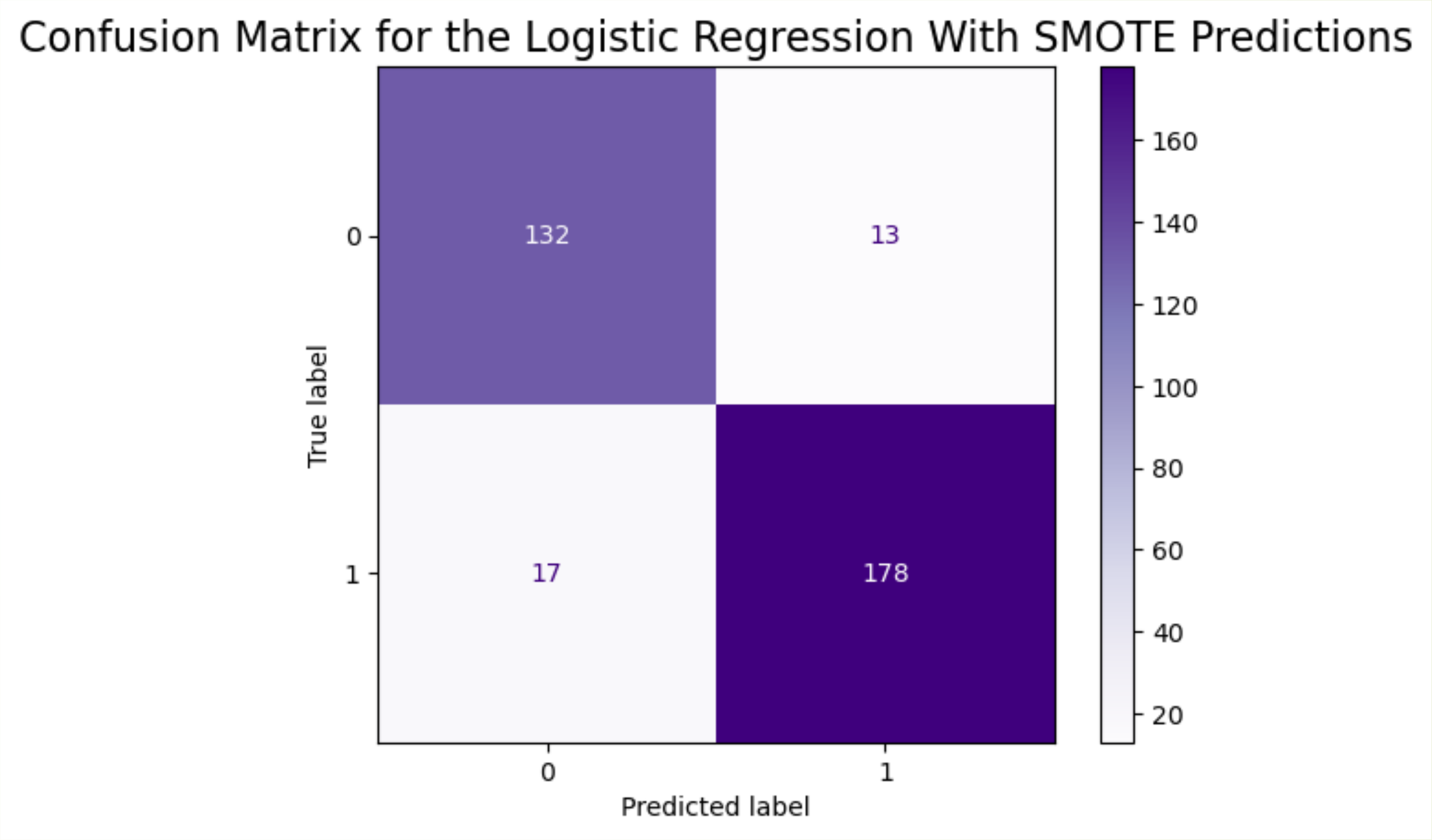
Classification Metrics (Continued)

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



Bonus: Logistic Regression with SMOTE

- Specificity | 0.91
- Sensitivity | 0.913
- Accuracy | 0.912
- Precision | 0.932
- Miscalculation Rate | 0.088



Confusion Matrix for the Logistic Regression With SMOTE Predictions

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)

Conclusion

- Optimal Model: TVEC with SVM
- Goal of the model: prioritize sensitivity, this result is achieved
- Positive class was whether the post belonged to the `democrats` subreddit class, denoted by a 1
- Precision was high: probability that the model is correct when it predicts an example to be in the positive class
- Accuracy, the percentage of observations correctly predicted within the test class, has peaked at 90%
- Here, *SVM* outperformed *logistic regression* in these aforementioned classification metrics.

Sources

- References from class work, lessons, and Reddit, PRAW, .sklearn documentation centers