Subreddit Classification Using Web APIs & NLP

Katie Sylvia - General Assembly - Project 3

Hello!

I am Katie Sylvia

Using NLP and classification techniques, can a model be created that outperforms our baseline model when predicting which of two subreddits a post came from?

Can this be done from the titles of two subreddits that focus on videos, gifs, and images rather than text?

Selected Subreddits

r/AnimalsBeingBros

“A place for sharing videos, gifs, and images of animals being bros.”

4,288,897 users

6500 posts

Collected from the beginning of every month from January 2016 to June 2021

r/AnimalsBeingJerks

“A place for sharing videos, gifs, and images of animals being jerks.”

4,032,329 users

6500 posts

Collected from the beginning of every month from January 2016 to June 2021

r/AnimalsBeingBros - Top of All Time -

88.7k Upvotes

TITLE: “These ten ducklings were found orphaned and they were brought to a pet duck called Stella who had just hatched nine of her own two weeks prior. She immediately claimed the ten as her own.”

r/AnimalsBeingJerks
- Top of All Time -

89.4k Upvotes

TITLE: “He would if he could”

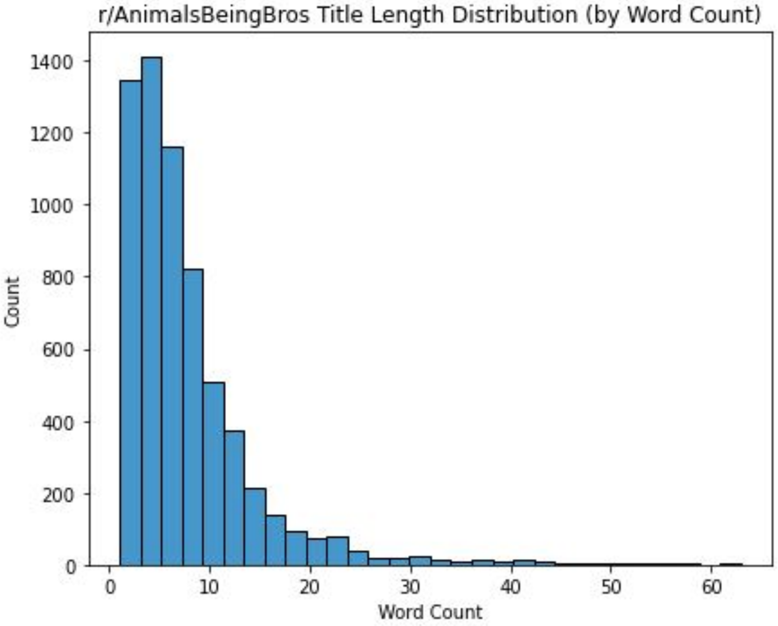
EDA & Pre-Processing

1. Title Length

What is the average title length for each subreddit?

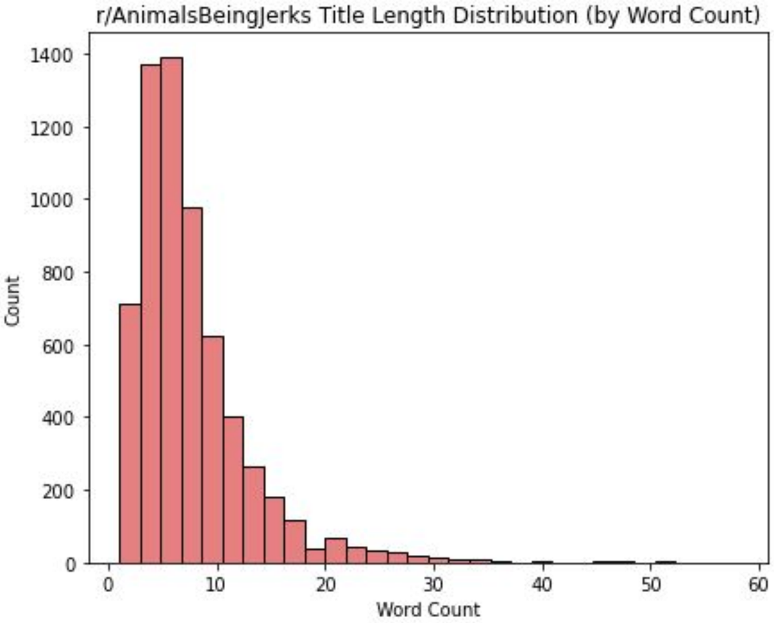
r/AnimalsBeingBros Titles

Mean Word Count: 8.0 words
Median Word Count: 6 words



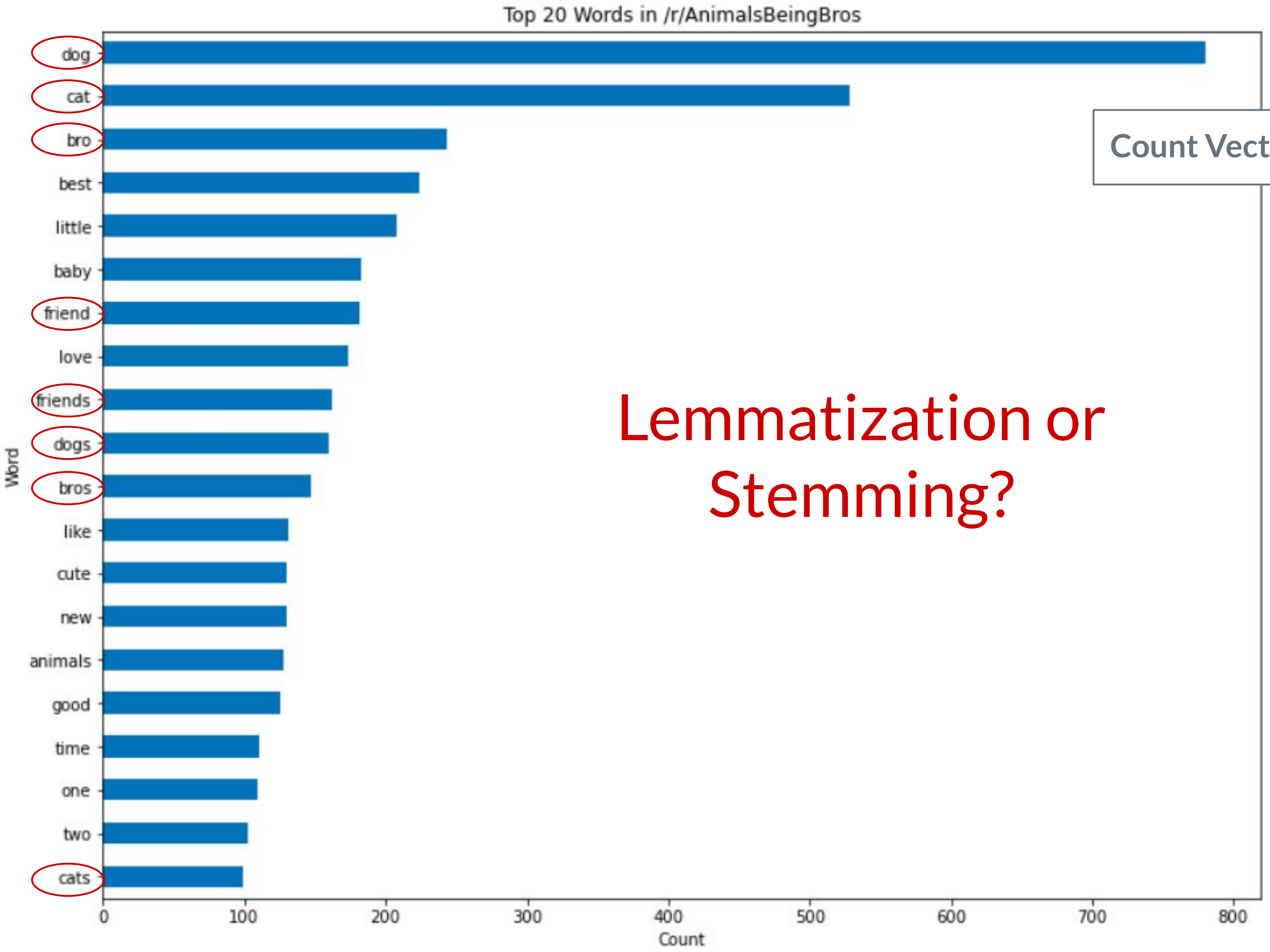
r/AnimalsBeingJerks Titles

Mean Word Count: 7.5 words
Median Word Count: 6 words



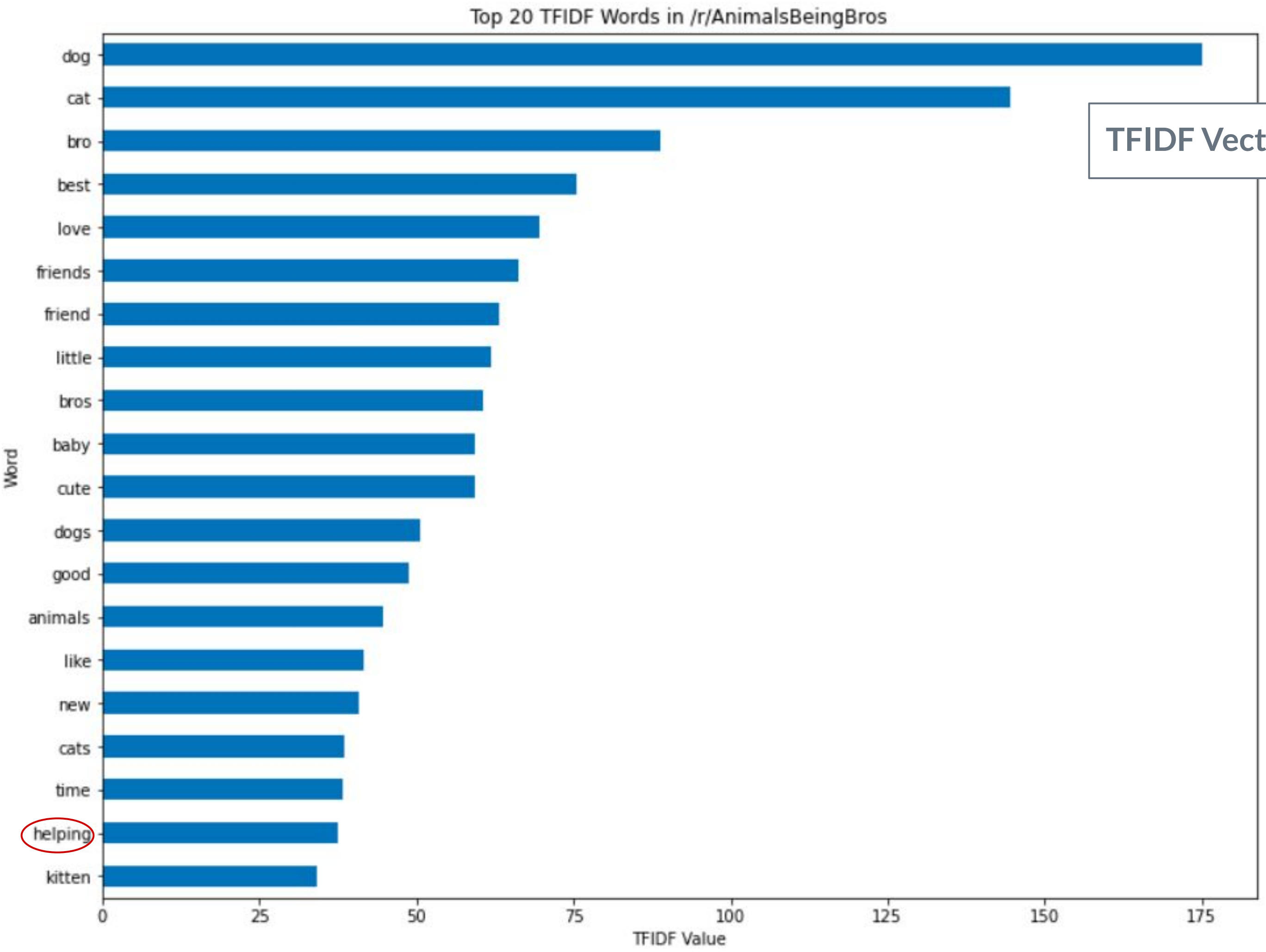
2. Most Common Words

What are the 20 most common words used in each subreddit?

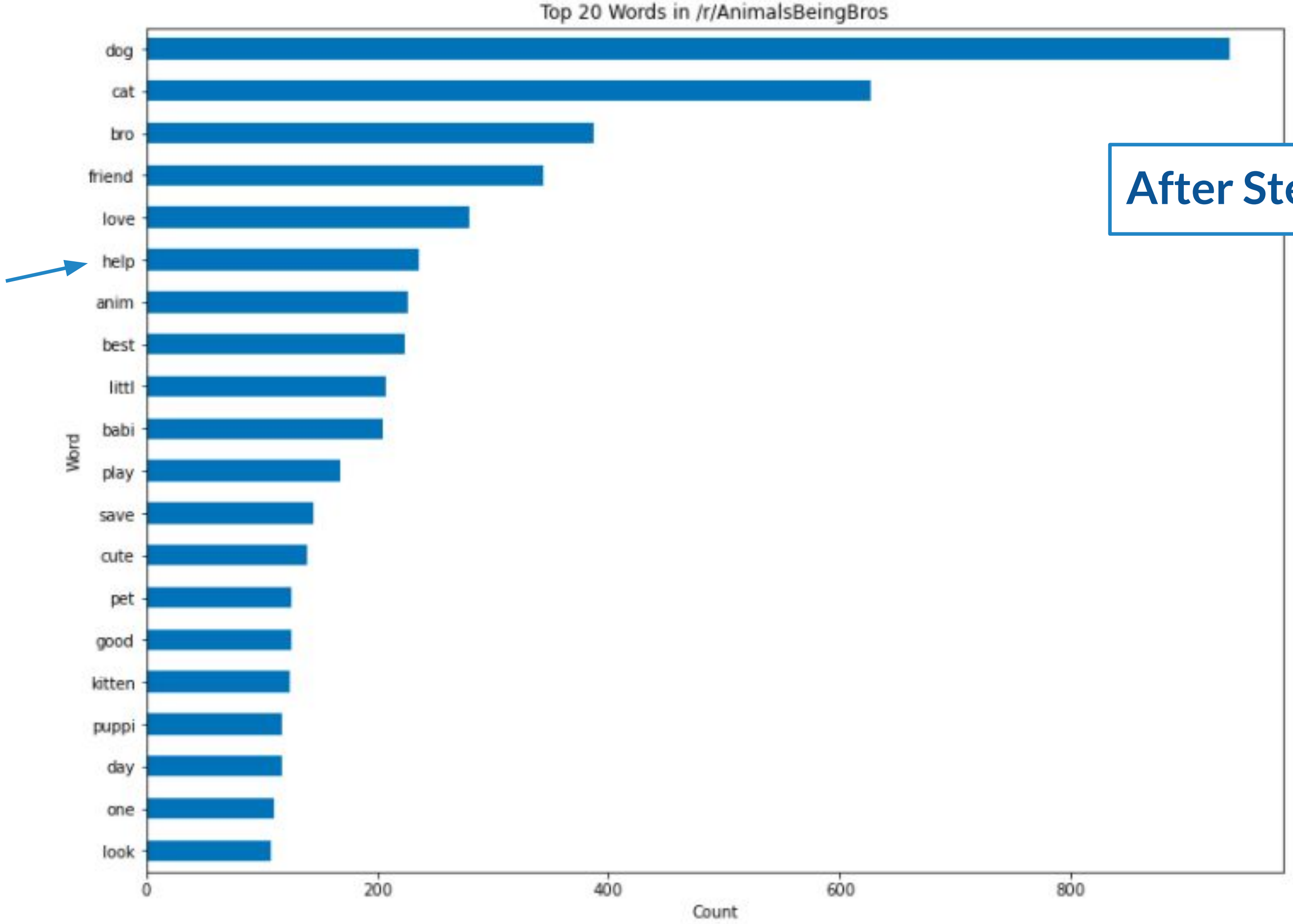


Count Vectorized

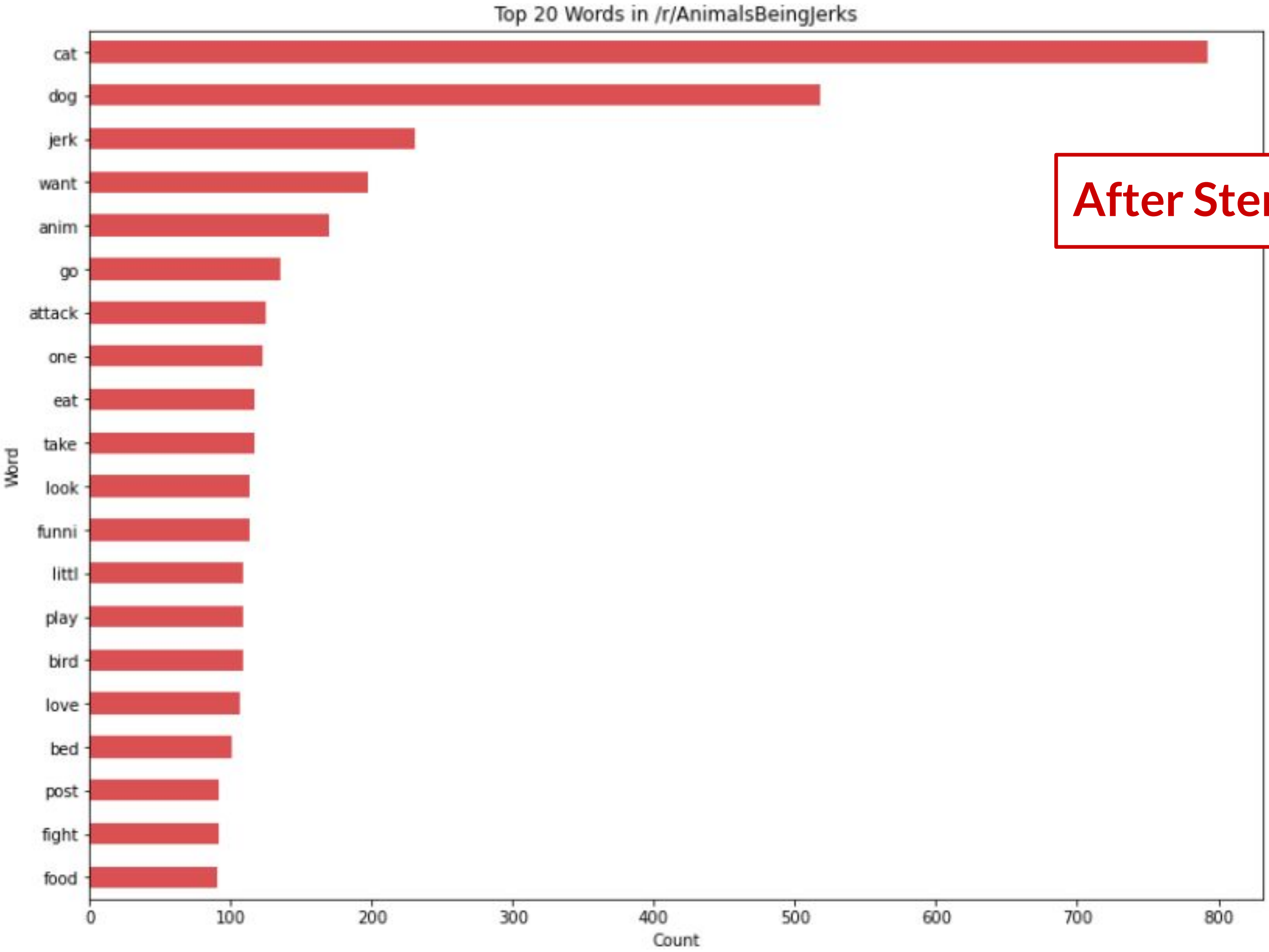
Lemmatization or Stemming?



TFIDF Vectorized



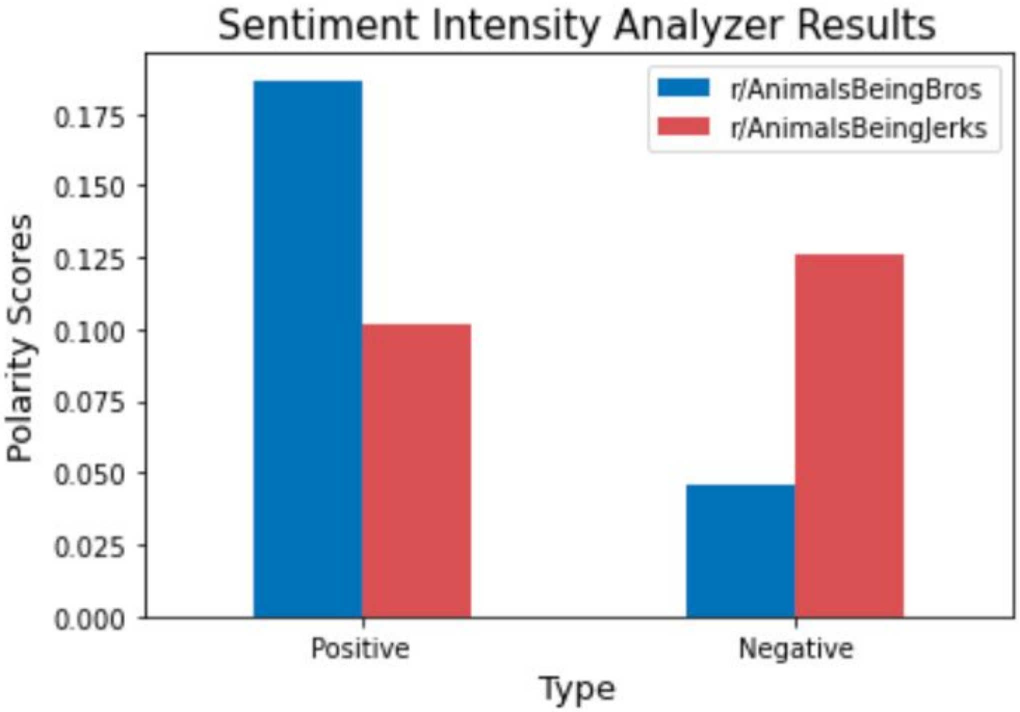
After Stemming



After Stemming

3. Sentiment Intensity Analyzer

What are the average positive, negative, and neutral polarity scores for each title in each subreddit?

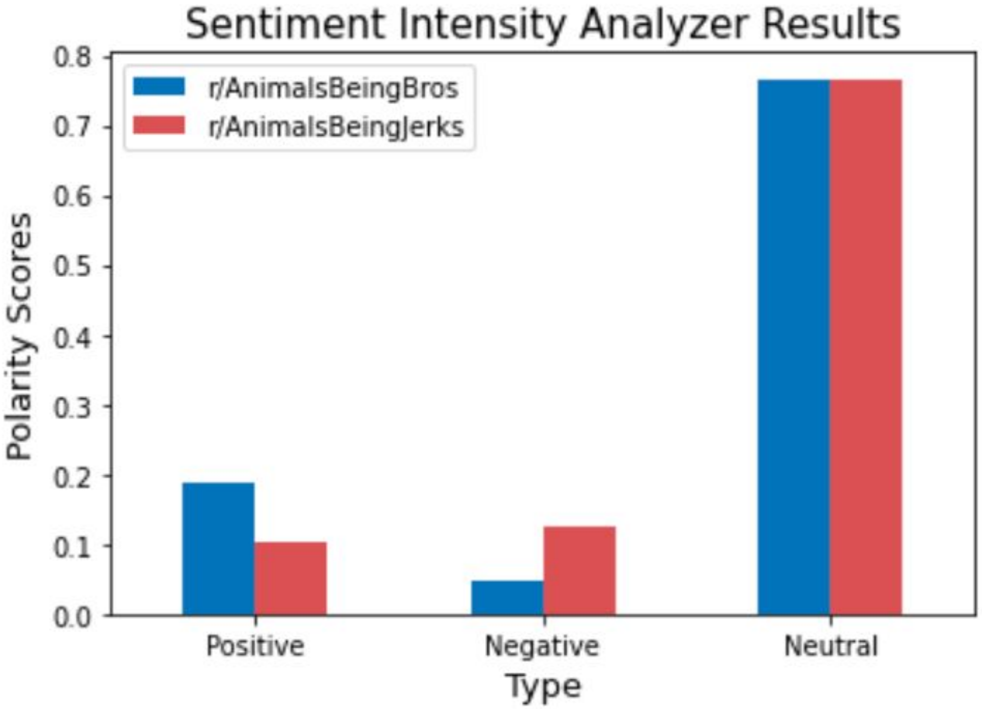


Positive Polarity Scores:

r/AnimalsBeingBros: 18.7%
r/AnimalsBeingJerks: 10.1%

Negative Polarity Scores:

r/AnimalsBeingBros: 4.6%
r/AnimalsBeingJerks: 12.6%



Though most of the text in titles from both subreddits were considered neutral.

Modeling &
Model Optimization

Baseline - 50.3%

Without using any NLP or classification techniques, my model would correctly predict a post belonged to r/AnimalsBeingJerks 50.3% of the time.

Models Tested

Logistic Regression

Bernouli Naive Bayes

Multinomial Naive Bayes

Random Forest

Gradient Boost

Ada Boost

Support Vector Machine

K-Nearest Neighbors

Pipeline and GridSearchCV tools were used to optimize the highest-scoring result.

Multiple pipelines were created for each model to run to assess the performance of:

- Stemmed titles vs. lemmatized titles
- TFIDF Vectorizer vs. Count Vectorizer

Once the best model was determined, hyperparameter tuning continued to optimize our model.

Best Model

Multinomial Naive Bayes

(Using stemmed titles and Count Vectorizer)

Best Parameters:

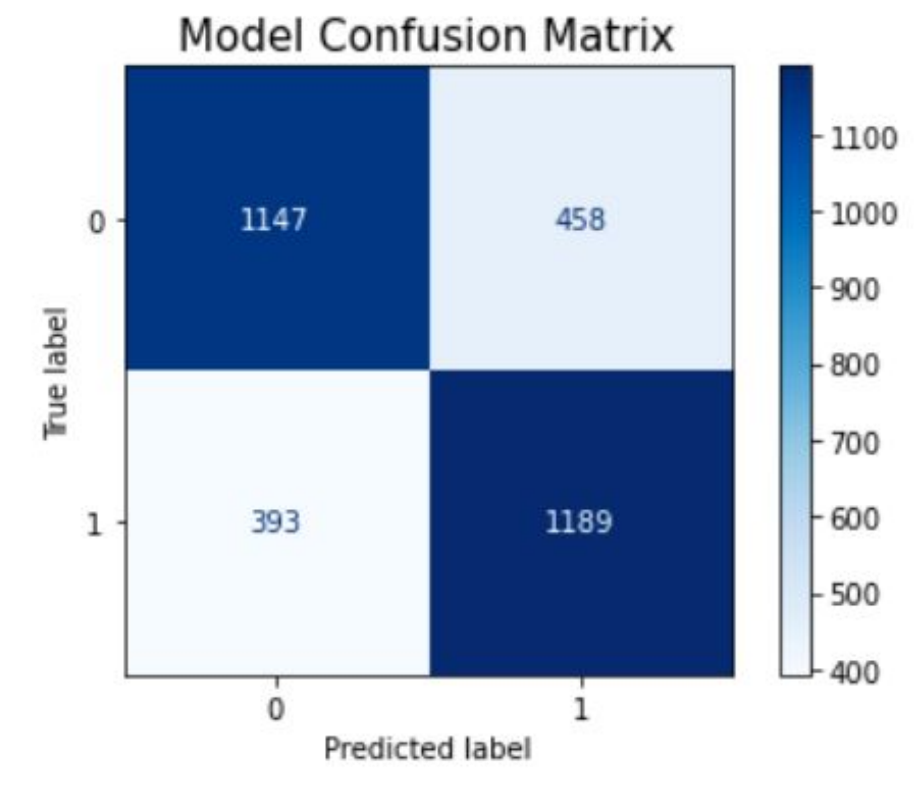
- cv__stop_words: None
- cv__ngram_range: (1, 2)

Scores

Training Score: 0.944

Testing Score: 0.733

Model Evaluation



This confusion matrix shows that the model incorrectly predicted r/AnimalsBeingJerks (1) 458 times when the post actually belonged to r/AnimalsBeingBros (0). Also, the model incorrectly predicted r/AnimalsBeingBros (0) 393 times when the post actually belonged to r/AnimalsBeingJerks (1).

Accuracy	73.3%
Sensitivity	75.2%
Specificity	71.5%
Precision	72.2%

Example of Title with High Probability (>99.99%) of Being in r/AnimalsBeingBros:

“Hero german shepherd shot multiple times saved his 16-year-old owner from burglar.”

Example of Title with High Probability (>99.99%) of Being in r/AnimalsBeingJerks:

“One of the cats threw up on my charging cord. It dried and I didn't notice when I put it into my phone to charge before passing out. Now my phones charging port is rusty. Dude ive had this phone for 4 months, my phone is pivotal in my art work.”

Examples of Misclassified Titles:

“Come at me bro.”

- r/AnimalsBeingJerks

“Choooommmmppppppppppppppppp”

- r/AnimalsBeingJerks

“😋”

-r/AnimalsBeingBros

“Stubborn english bulldog refuses to get off bed.”

- r/AnimalsBeingBros

“arrrrgh”

- r/AnimalsBeingBros

“My best friend forever.”

- r/AnimalsBeingJerks

Recomendations & Conclusions

Recommendations to Improve Our Model

Collect More Data

Since both of these subreddits are not text heavy, more posts should be collected.

Include Comments

At just an average of 6 words in each title, adding comments to our data would likely improve our model.

Investigate More Parameters

As wonderful as Pipelines and GridSearchCV are, using them in practice can be incredibly time consuming. More extensive searches with different parameters can be conducted to further optimize our model.

With an average of just 6 words in each title to make predictions on, the model performed well.

The question remains answered; with a 45.7% improvement from our baseline model, yes, a model can outperform our baseline model when predicting which of two non text-heavy subreddits a post came from.

Thanks!

Any questions?