Solving a Reddit Problem

By Christopher Kuzemka

A Poorly Merged Subreddit Exists

My girlfriend loves to use Reddit. One of her favorite subreddits is "r/aww", a community dedicated forum largely consisting of cute animals and cute moments captured on video and on camera. However, there is another reddit that is the complete opposite of cute animals and cute moments captured on video and on camera -- this subreddit is known as "r/natureismetal" -- and it was merged together with "r/aww" to create a "super-subreddit" known as "r/dangerouslycute." This made my girlfriend very upset, as she was never a fan of the content from "r/natureismetal" and now she is tainted by its controversial content. We can also imagine that many others must also feel the same way about the merge.

https://www.wikihow.com/Comfort-Your-Girlfriend-when-She-Is-Upset

Using data collected previously from the subreddits before the merge, we are going to utilize Natural Language Processing classification models to separate the subreddit content. Our supervised learning models will be judged by their accuracy measure for success. The models we will explore will be LogisticRegression, Multinomial Naive Bayes, and Gaussian Naive Bayes with a use of CountVectorizer and TFIDFVectorizer across our data. We will do an in depth analysis on a successful model and explore the various quirks behind the influences of its predictions.

https://i.4pcdn.org/pol/1504497858143.gif

What are our Target Subreddits?

The member size of r/aww is 24,469,289 users

r/aww

https://www.reddit.com/r/aww/comments/g6ybfo/cute_doggo/

r/aww's Summary:

- Happy content
- Cute pictures
- No NSFW content
- No death related content

r/natureismetal's Summary:

- Graphic content
- Intimidating content
- NSFW content allowed

The member size of r/natureismetal is 1,469,418 users

r/natureismetal

https://www.reddit.com/r/natureismetal/comments/g4s0f0/not_a_happy_lion/

1) What ACCESS Do We Have To The Data?

2) What DATA Do We Have Access To?

Number of Comments Per Subreddit

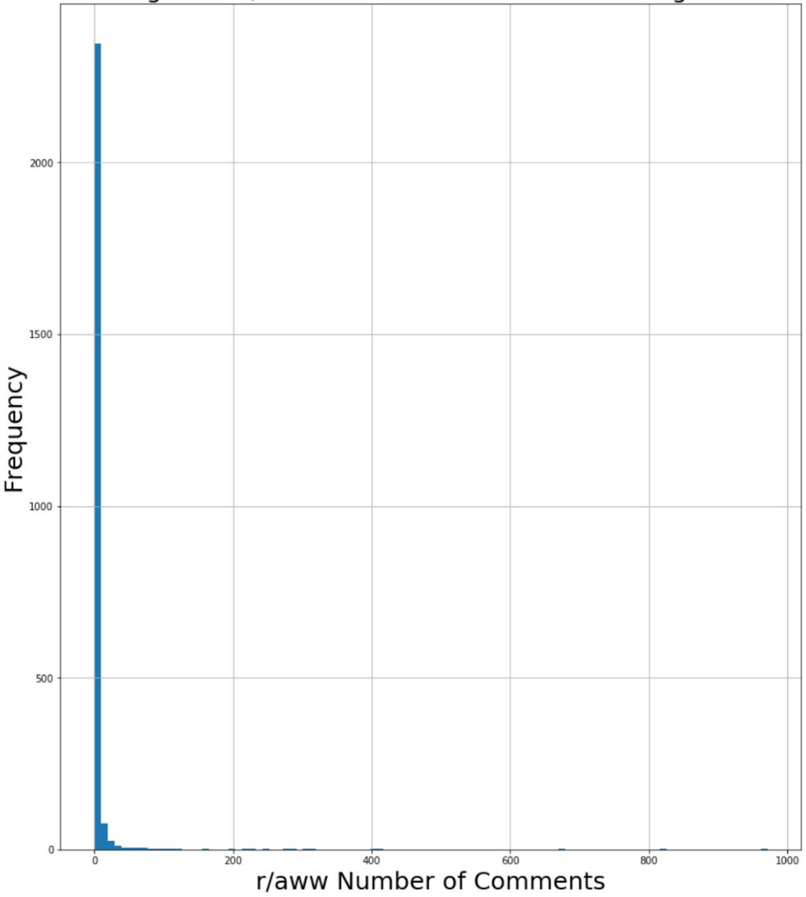


Figure 1: r/aww Submission Comments Histogram

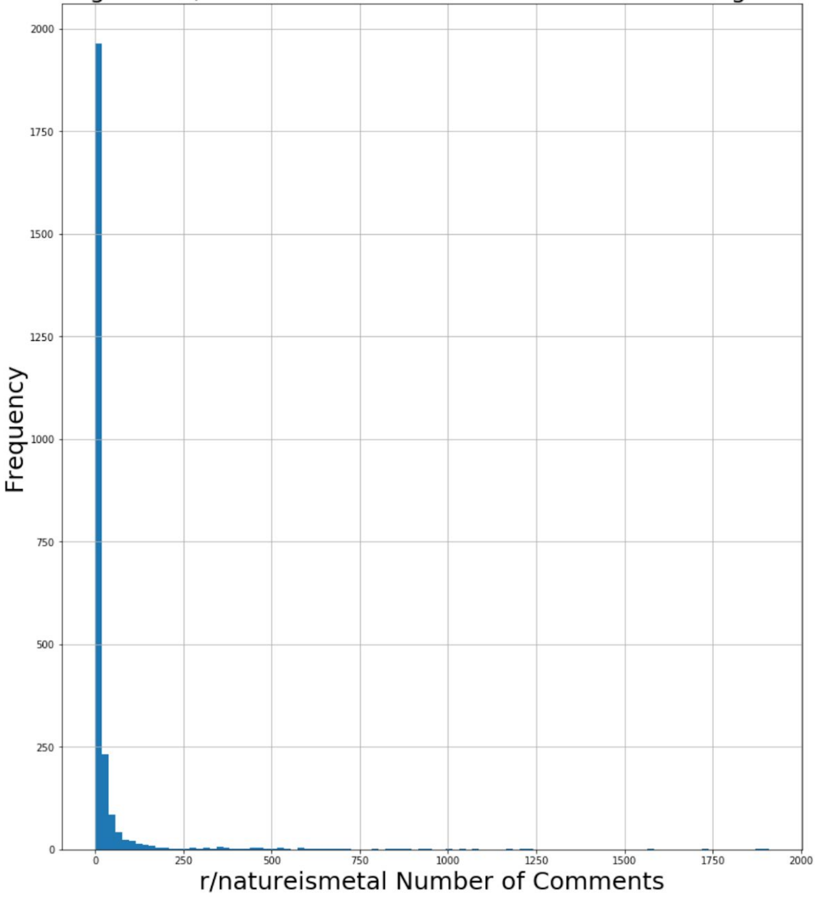


Figure 2: r/natureismetal Submission Comments Histogram

20 Most Common Words per Subreddit (With Custom English Stopwords)

Figure 6: r/aww 20 Most Common Words in Submission Titles

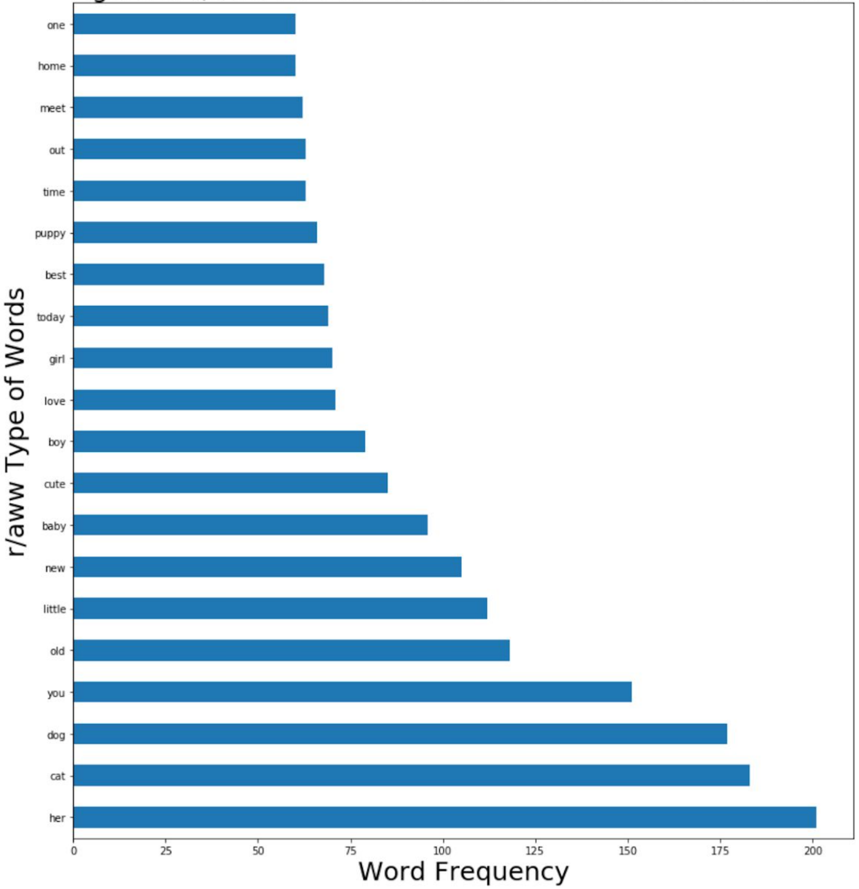
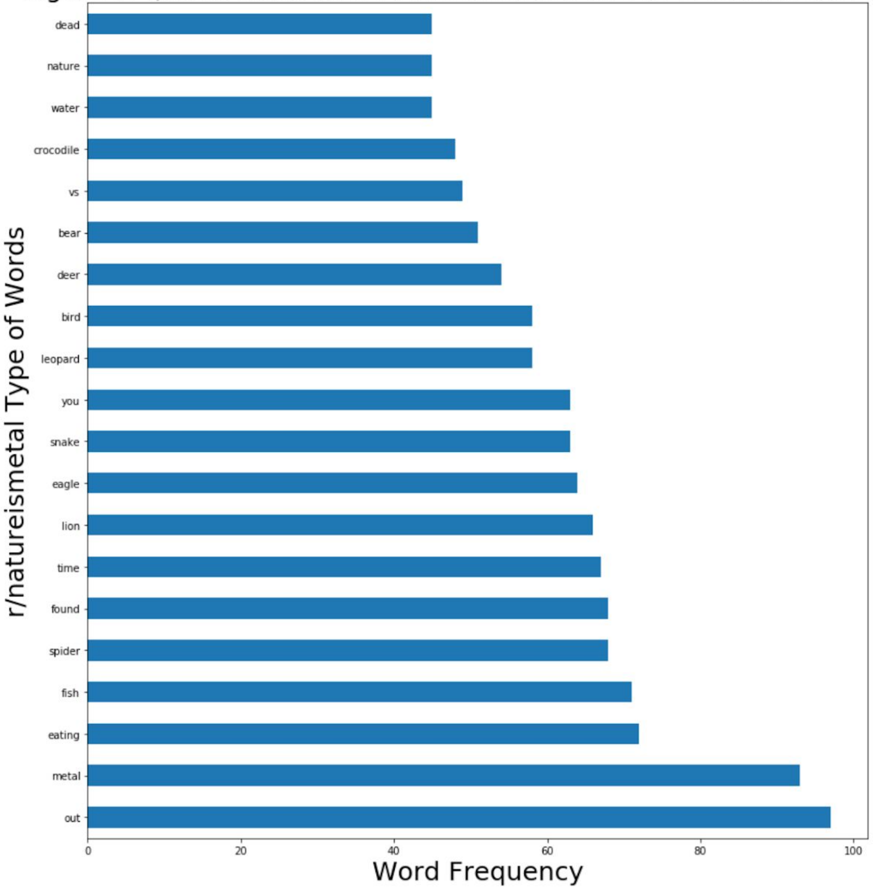
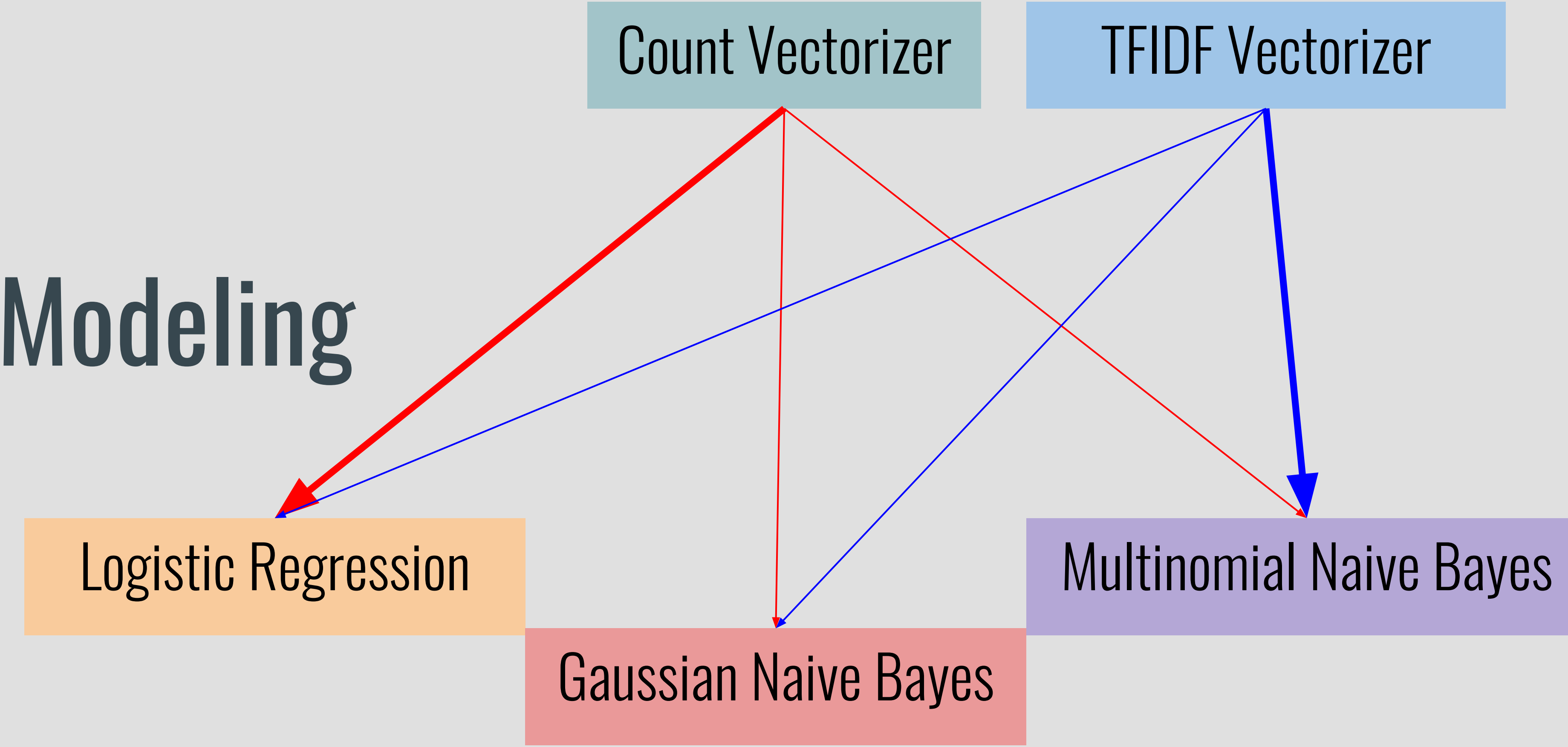


Figure 7: r/natureismetal 20 Most Common Words in Submission Titles



Modeling



Count Vectorizer

TFIDF Vectorizer

Logistic Regression

Gaussian Naive Bayes

Multinomial Naive Bayes

Logistic Regression & Count Vectorizer

Key Interpretable Model Findings (at 83% accuracy):

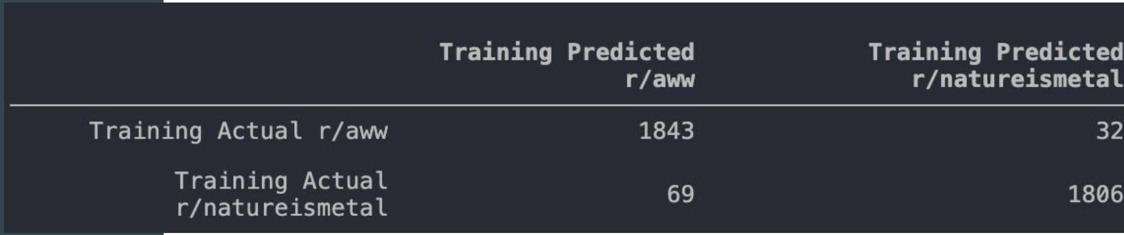
Use of the word “frogger” meant you are ~3.92 times more likely to be affiliated with r/aww, all else held equal.

Use of the expression “giant horned” meant you are only ~0.16 times likely to be affiliated with r/aww, all else held equal.

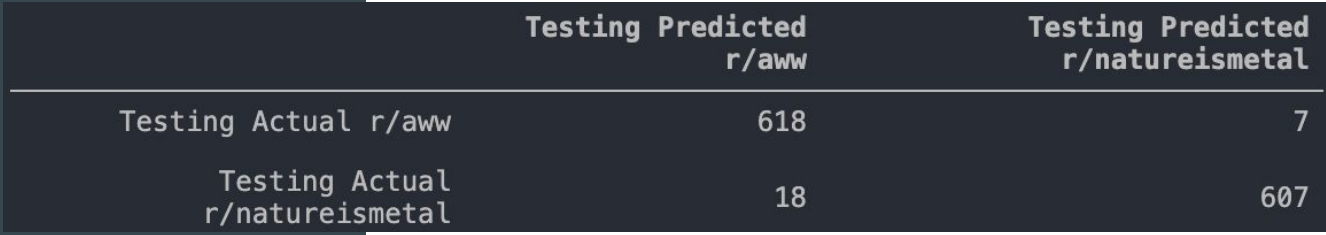
Multinomial Naive Bayes & TFIDF Vectorizer

Training Accuracy: 89%

Testing Accuracy: 84%



	Training Predicted r/aww	Training Predicted r/natureismetal
Training Actual r/aww	1843	32
Training Actual r/natureismetal	69	1806



	Testing Predicted r/aww	Testing Predicted r/natureismetal
Testing Actual r/aww	618	7
Testing Actual r/natureismetal	18	607

Conclusions

- We need more time, data, and resources to get a fully functioning model that will be the backbone for our application.
- Within reason and understanding of error due to data bias, you can separate the subreddit content with upwards of 84% accuracy.
- Should consider ensemble modeling methods for future. Use of servers to be included to speed calculations.
- Should consider pulling in more features.