

# Dog Reddit Analysis

by Tonya Chernyshova

# Goal

To create a model which can identify dog lovers and dog haters based on their posts on different social networks and websites using NLP.

*This information could be used further for targeting ads.*


# Subreddits picked

Dog haters

VS.

Dog lovers

**COMMUNITY DETAILS**


 **r/Dogfree**

**14.7k**  
Members

**128**  
Online

Not everyone loves dogs

**COMMUNITY DETAILS**

 **r/dogs**

**1.0m**  
Members

**887**  
Online

/r/dogs is a discussion-based subreddit, meant for asking questions, sharing information, and learning about our beloved canine companions and related dog-centric topics.

[illegible]

# Methodology

## Collecting and Cleaning Data:

- Gathering data from Reddit's API using AWS and Cron
- Cleaning data (punctuation, capitalization)
- Combining title and post's text, stemming data
- Binarizing target variable

# Methodology

## Preprocessing and Modeling Data:

- Split data into train and test (75%/25%)
- Grid searched 3 models using TF-IDF and Count Vectorizers
  - Logistic Regression*
  - Multinomial Naive Bayes*
  - Random Forest Classifier*
- Evaluated models
- Chose the best model

# Results: compare models

|                   | Model 1:<br><u>Logistic Regression</u> | Model 2:<br><u>Naive Bayes</u> | Model 3:<br><u>Random Forest</u> |
|-------------------|--|--------------------------------|----------------------------------|
| Training Accuracy | 0.9832                                 | 0.9250                         | 1.0                              |
| Testing Accuracy  | 0.9519                                 | 0.9278                         | 0.9345                           |
| Overfit?          | slightly                               | no                             | yes                              |

# Results: best word predictors

Lovers: party, puppy, breed, food

Haters: Attack, owners, shit



# Conclusions

- The best model is Logistic Regression, we can predict with an accuracy of about 95% where a given post came from, 19 out of 20 posts.
- Company can deliver relevant ads to Dogs' lovers and avoid Dogs' haters
- All dogs are full & happy

