

# Reddit Posts Analysis with Natural Language Processing

By Daniel Seto, Florian Combelles, Goh Yan Da,  
Joanne Chong, Kenneth Goh





# Table of Contents



**Executive Summary**



**Problem Statement**



**Methodology**



**Data Collection**



**Exploratory Data Analysis**



**Classification Models**



**Conclusion & Recommendations**

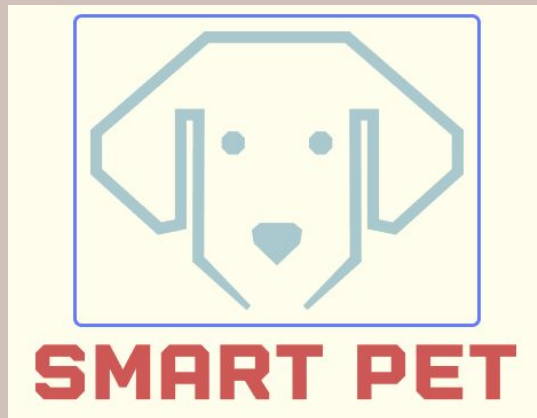
1

# Executive Summary



# Who we are

- **Smart Pet**, established in **August 2017** by *Florian Combelles*
- Passion for animals with expertise from former vets, pet store owners and animal shelter volunteers
- Worked with organisations and companies such as:
  - **SPCA**
  - **AVA**
  - **Petslovers**



*Our mission:*  
*"We care for them"*





2

# Problem Statement

The challenge at hand

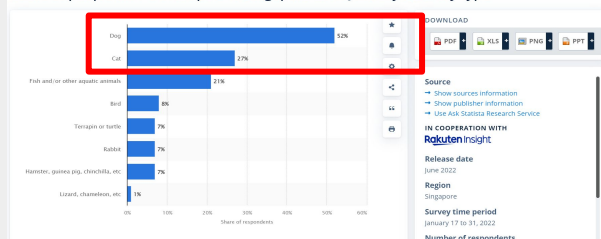


# Background

- Most common pets in Singapore
  - **Dogs - 52%**
  - **Cats - 27%**

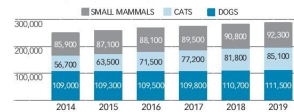
↑ Increased interest in pet ownership → Rise of inexperienced pet owners

Share of people who owned pets in Singapore as of January 2022, by type



## Growing pet population

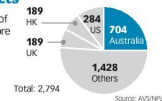
Singapore's population of pet dogs, cats and small mammals like hamsters, guinea pigs and rabbits has been growing over the years



Note: Figures include pets not registered with AVIS as well as adopted pets  
Source: Euromonitor International

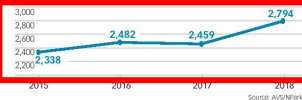
## Top import markets

Of the 2,794 total imports of dogs and cats into Singapore in 2018, about half were sourced from just four markets: Australia, the US, Hong Kong and the UK.



## Foreign furry friends

Total net imports into Singapore



**THE STRAITS TIMES** **LIFE**

More people in Singapore interested in adopting or fostering pets during Covid-19 pandemic

**LIFE**

Centre, says that cut off from social contact, people feel isolated. Pets, especially dogs and cats, make wonderful companions, providing warmth and distraction, she adds.

Ms Jiny Mohandas Special, in May. PHOTO: COURTESY OF JINY MOHANDAS

At Causes for Animals, co-founder Christine Bernadette, 31, notes an increase in adoptions during the pandemic.

Before April, the organisation had six to eight dog adoptions and two cat adoptions a month. Since the circuit breaker, the number has jumped to 10 to 12 dogs and five cats a month.

At Chained Dog Awareness In Singapore, a volunteer-run advocacy group which specialises in helping dogs suffering from confinement or tethering for long hours, co-founder Lee Pin observes a spike in the number of people looking to adopt or foster dogs since phase two, which started on June 19.

In June, July and this month, the group received 20 such enquiries - double the number pre-Covid-19.

Mr Colin Chew, 52, a volunteer with Just For Paws, says enquiries for adoption doubled during the circuit breaker, compared with the same period last year. Since



With an influx of **inexperienced** pets owners  
**overly reliant** on vets and pet store, how can  
we better optimize everyone's time?

# Stakeholders' Concerns

- Lack of local resources regarding information on:
  - Licensing, upkeep costs, aftercare

## How does it affect you?



Decrease in work efficiency

- Pet store owners occupied answering queries



Overbooking of Vet appointments

- Pet owners coming in for minor enquiries/non-emergencies



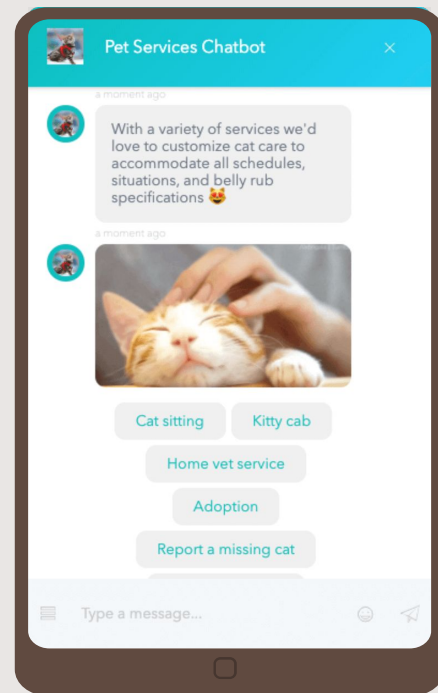
More animals abandoned

- Animals abandoned after pandemic are sent to the shelters



# Our mobile app (beta)

- **Pet Companions**, mobile application (currently in beta) that can offer pet owners information through:
  - AI-powered chatbot
    - First level of customer inquiry
    - Source of information for pet owners
  - Recommends articles based on classification of queries
  - A knowledge-bank filled with information sourced from vets, pet experts and users



3

# Methodology

How we've been doing it



# Existing model - K-Nearest Neighbours (KNN)

**78%**  
Accuracy

How does this work?

Kennel	
Litter box	
Resource guarding	
Long walk	
Meowing	
Scratching post	

- Classification of observation depends on the number of surrounding data points taken as reference

E.g. Classification of 'Resource guarding':

- If  $K = 3 \rightarrow$  Dog
- If  $K = 5 \rightarrow$  Cat

# Potential model - Naive Bayes



P (dog-related post): 0.60

15	10	4
Toy	Walk	Night

10	2	9
Toy	Walk	Night



P (cat-related post): 0.40

- Classification of the text post is based on comparing the conditional probabilities of it belonging to either r/CatAdvice or r/DogAdvice

“night walk”



- $P(\text{Cat} \mid \text{'night walk'})$  vs  $P(\text{Dog} \mid \text{'night walk'})$

# Potential model - Logistic Regression



- Each word/token within the model are weighted according to their importance to the respective categories
- The model will generate the probabilities for each category
- The text post will be labelled according to the category with the highest probability



4

# Data Collection & Cleaning

How we collect and clean the information





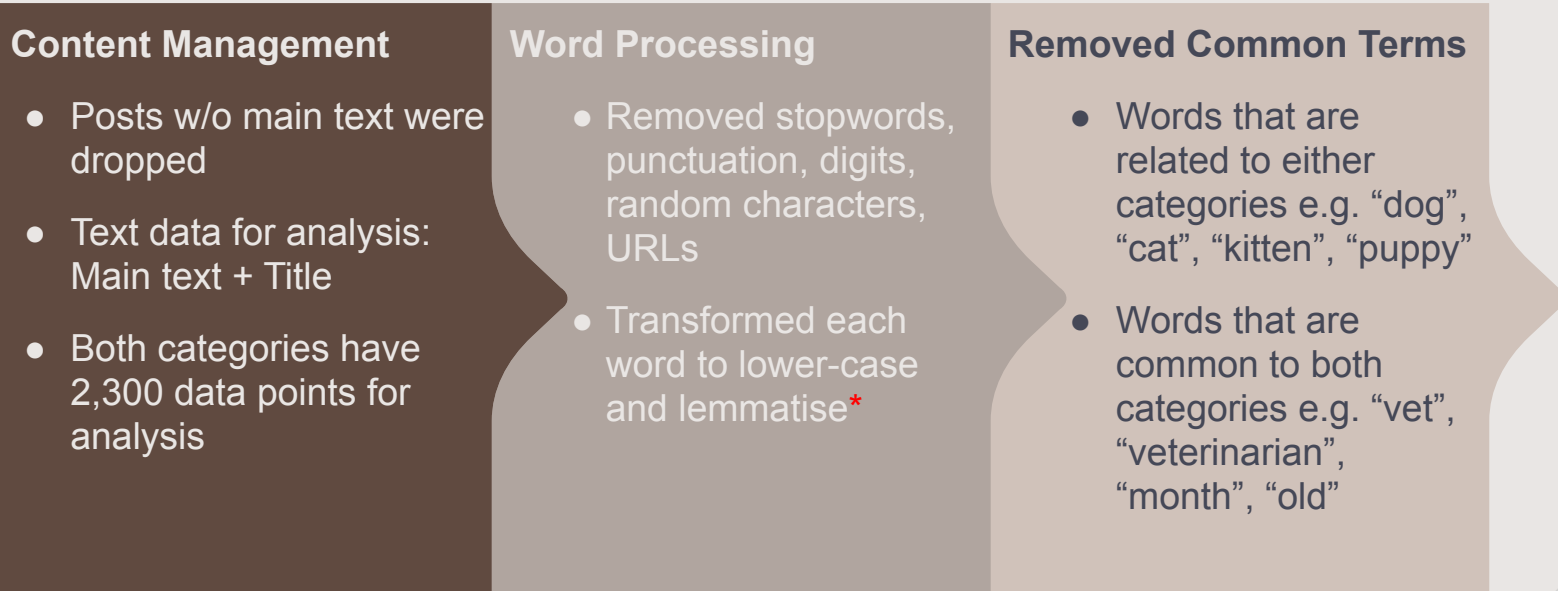
# Overview of data collected



- *r/CatAdvice* and *r/DogAdvice*
- User-generated data scrapped using Pushshift API

	<b>r/CatAdvice</b>	<b>r/DogAdvice</b>
<b>Period of Posts</b>	12 Oct - 24 Nov	25 Aug - 24 Nov
<b>No. of Posts</b>	4,247	4,000
<b>Size of Community</b>	119K	66.5K

# Cleaning process



\*Lemmatisation is the grouping of inflected forms of similar words so that they can be analysed as a single term (often as the root word)

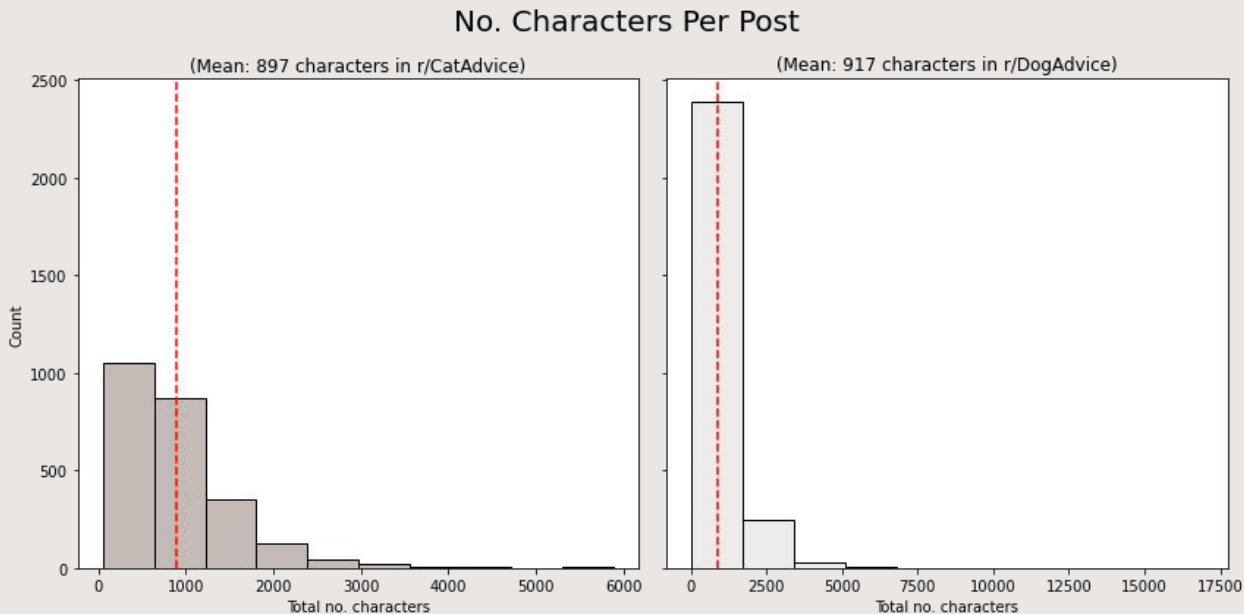
5

# Exploratory Data Analysis

Zooming in on our data

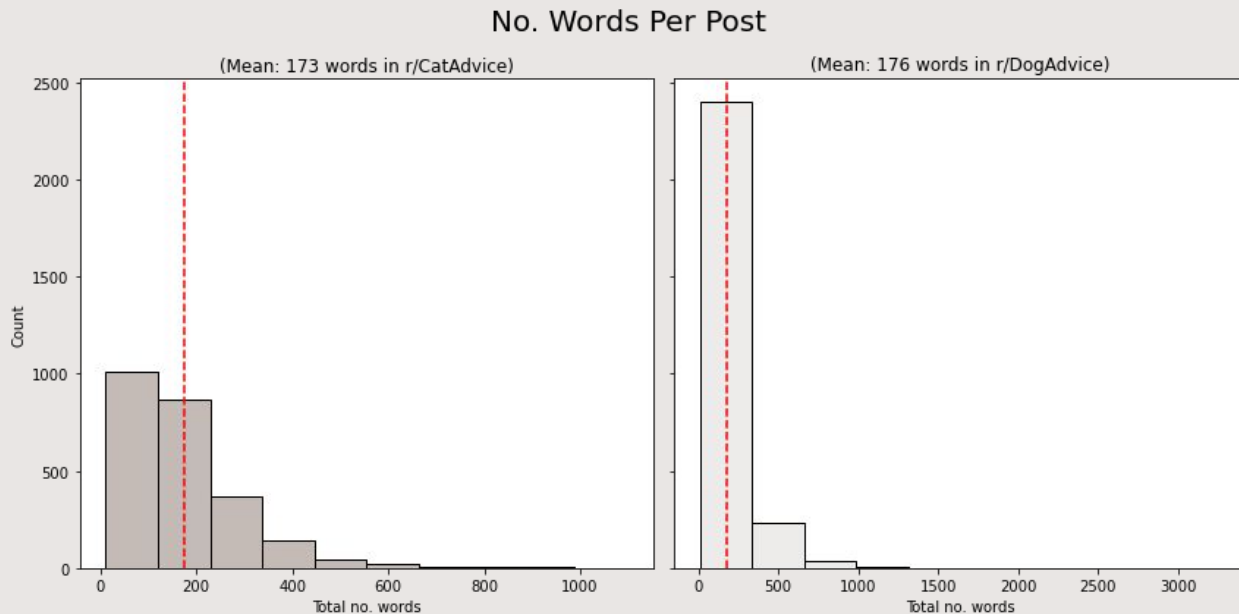


# Higher average character count in r/DogAdvice



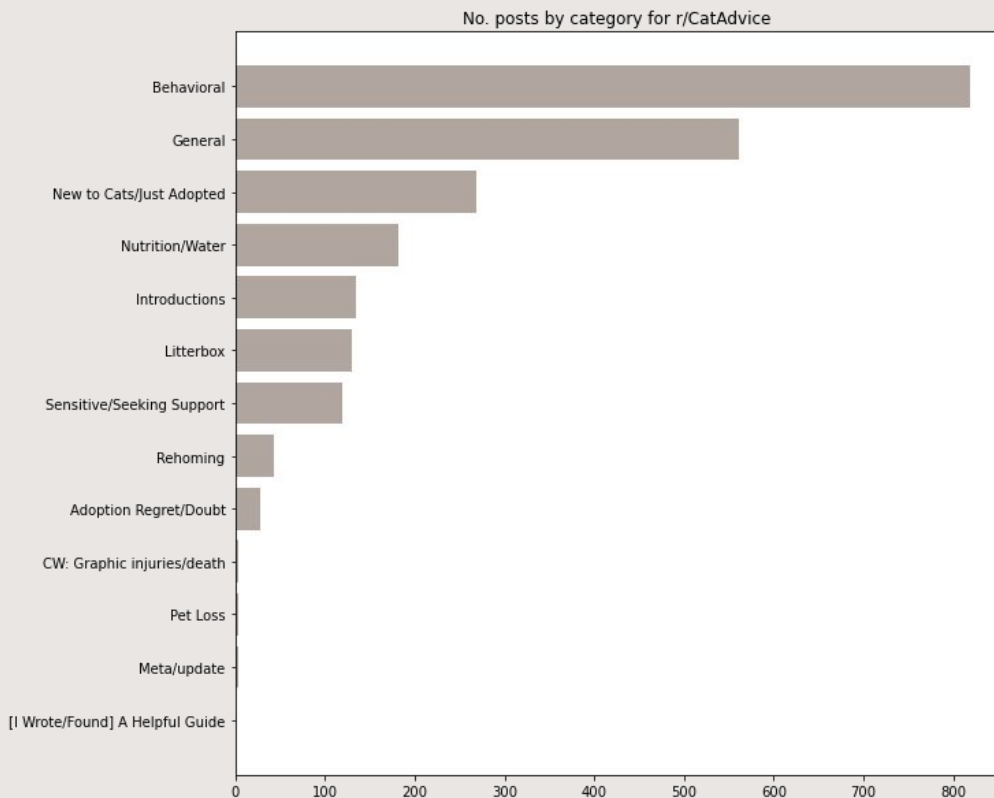
- **Maximum no. characters:** 5,885 in r/CatAdvice & 16,898 in r/DogAdvice
- **Minimum no. characters:** 56 in r/CatAdvice & 39 in r/DogAdvice

# Similar word count average in both subreddits



- **Maximum no. words:** 1,097 in r/CatAdvice & 3,267 in r/DogAdvice
- **Minimum no. words:** 11 in r/CatAdvice & 9 in r/DogAdvice

# Behavioral queries tops r/CatAdvice discussion

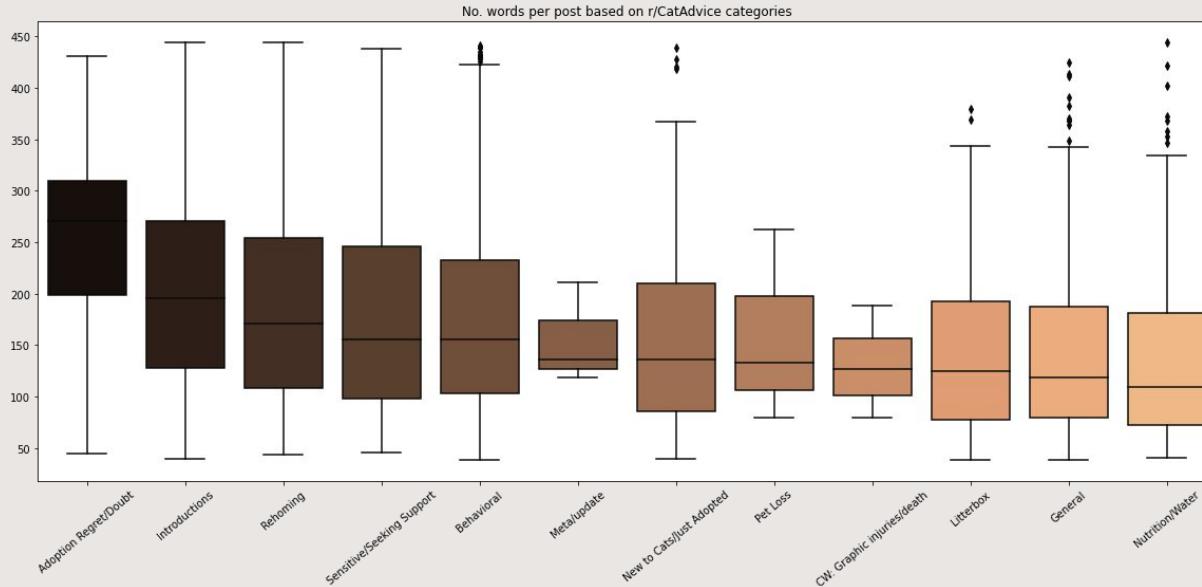


Top topics tagged:

- **Behavioral:** 818 posts
- **General:** 561 posts
- **New Cats/Newly adopted:** 269 posts

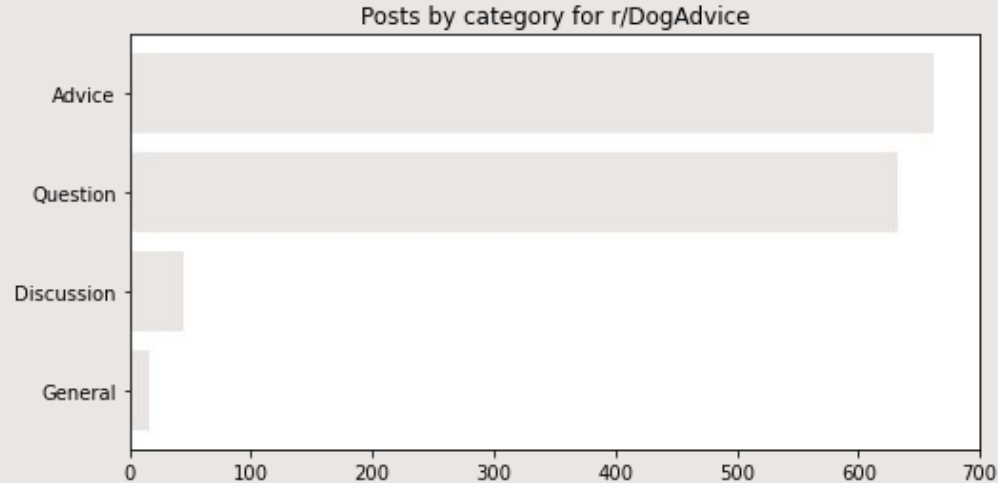


# Highest word count average on Adoption Regret/Doubt



- **Adoption Regret/Doubt:** 248 average no. words
- **Introductions:** 206 average no. words
- **Behavioral:** 175 average no. words

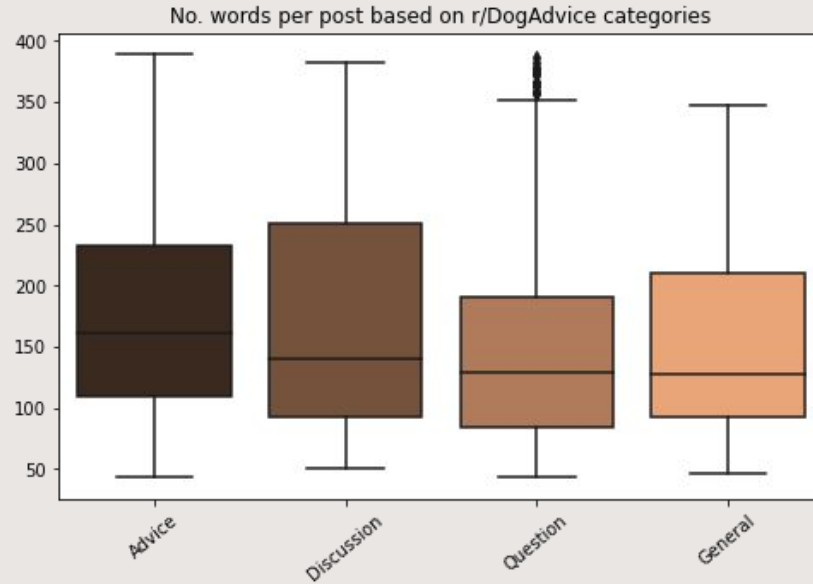
# More general categories in r/DogAdvice



Top topics tagged:

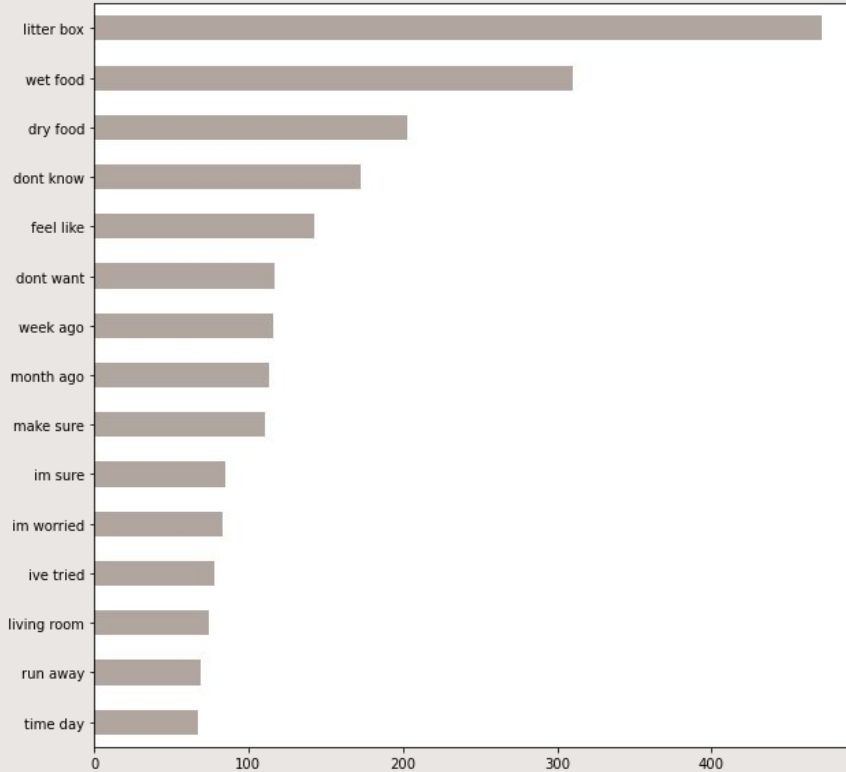
- **Advice:** 662 posts
- **Question:** 632 posts

# Highest word count average for Advice category



- **Advice:** 177 average no. words
- **Discussion:** 174 average no. words

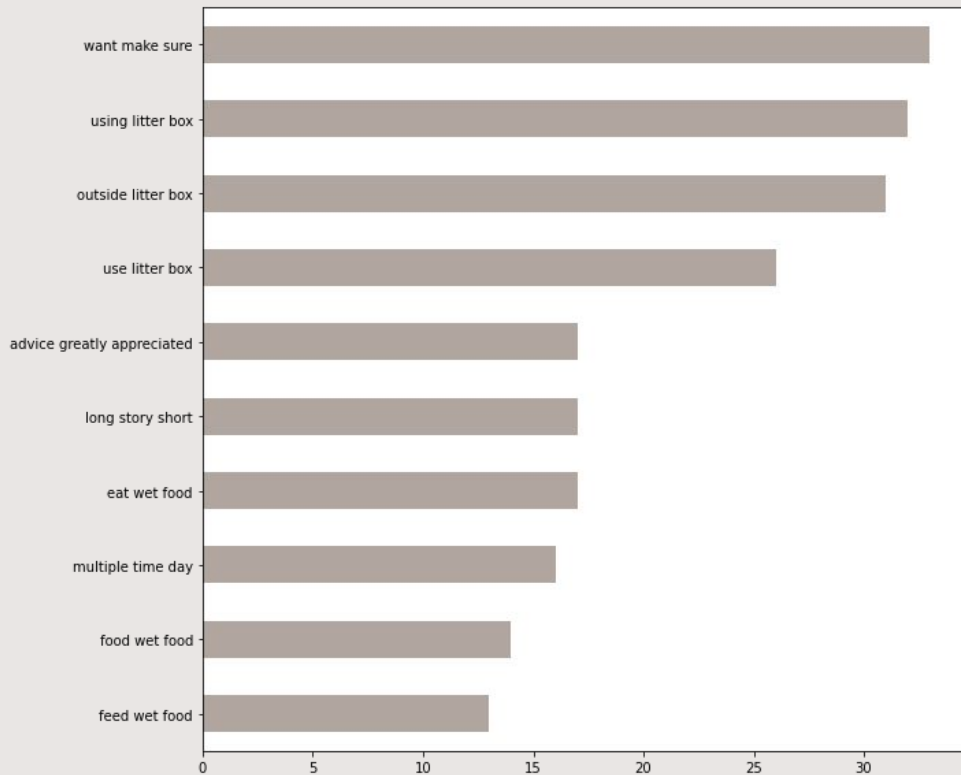
# Top 2-word queries in r/CatAdvice



- 'Litter box'
- 'Wet food'
- 'Dry food'

Users are concerned with their cats' environment and nutrition.

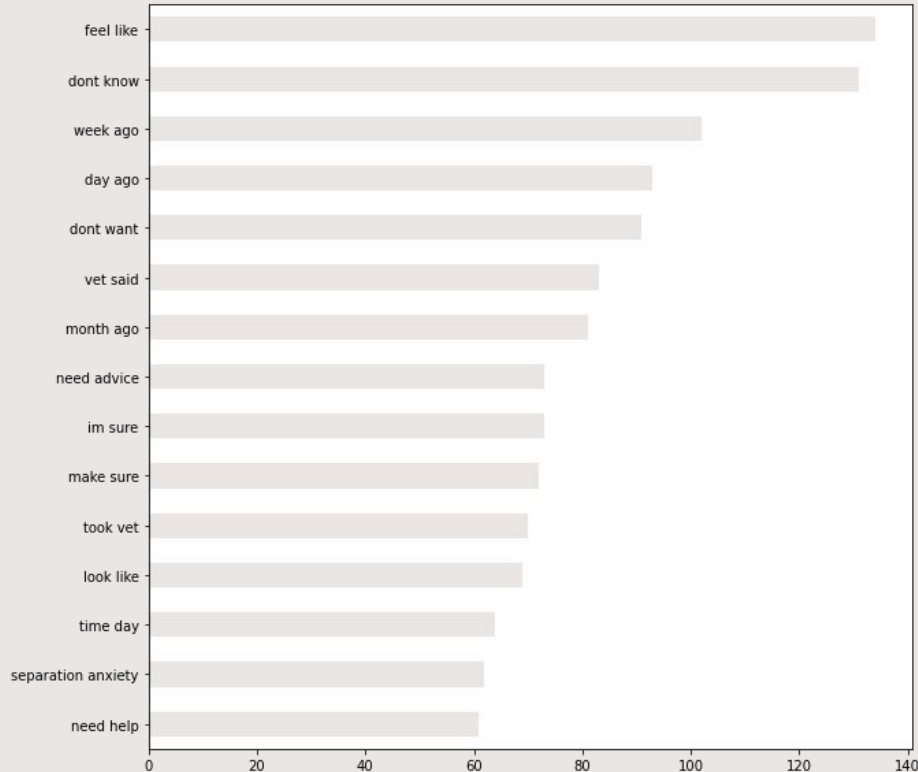
# Top 3-word queries in r/CatAdvice



- 'Using litter box'
- 'Outside litter box'
- 'Use litter box'

This reaffirms that users are most concerned with their cats' environment.

# Top 2-word queries in r/DogAdvice

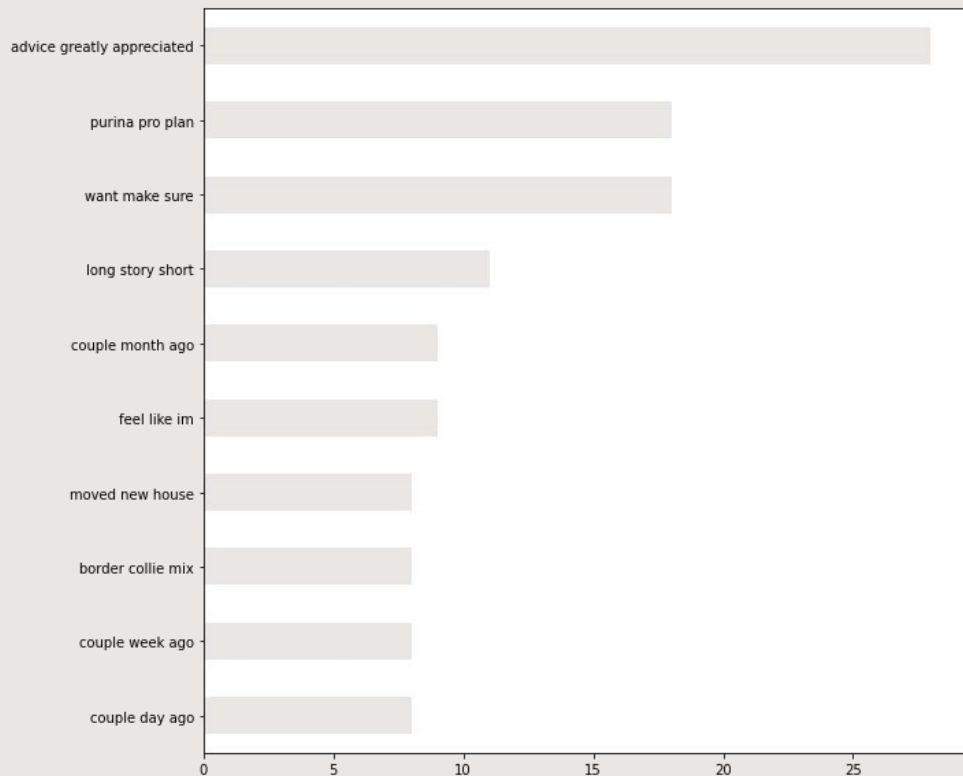


- 'Feel like'
- 'Don't know'
- 'Need advice / help'
- 'Separation anxiety'

Users seem mostly uncertain about the issues faced. The only issue mentioned is 'separation anxiety'.



# Top 3-word queries in r/DogAdvice



- 'Advice greatly appreciated'
- 'Purina pro plan'
- 'Want make sure'

This reaffirms users in this forum could be more uncertain and may require more support.

'Purina pro plan' was mentioned often - this is a formula to improve dog's digestion.

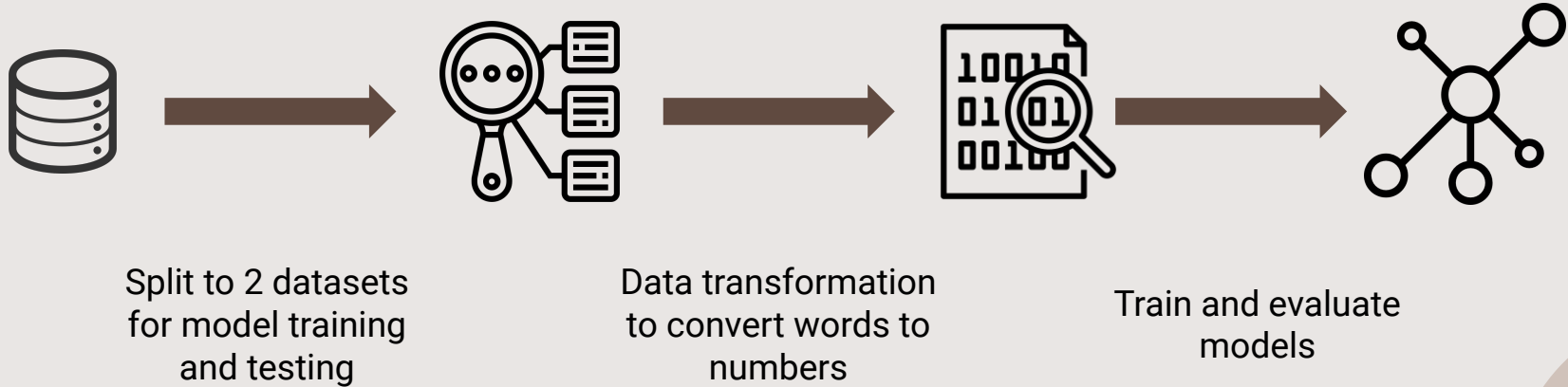
6

# Classification Models

Processing our data



# Data preparation and modelling



# The training models and measure of success

## Classification Models:

- Logistic Regression
- Naive Bayes (Multinomial, Bernoulli and Gaussian)

## Evaluation Metrics:

- Accuracy Score  
*No. of observations classified correctly /  
Total positive and negative observations*
- F1 score  
*Balance of precision and recall on positive  
observations*

## What we're looking for:

- ↑ Higher Accuracy Score
- ↑ Higher F1 score
- ↑ Better Model Performance

**Cat or dog?**



# Multi Naive Bayes achieved 90% accuracy & F1 score

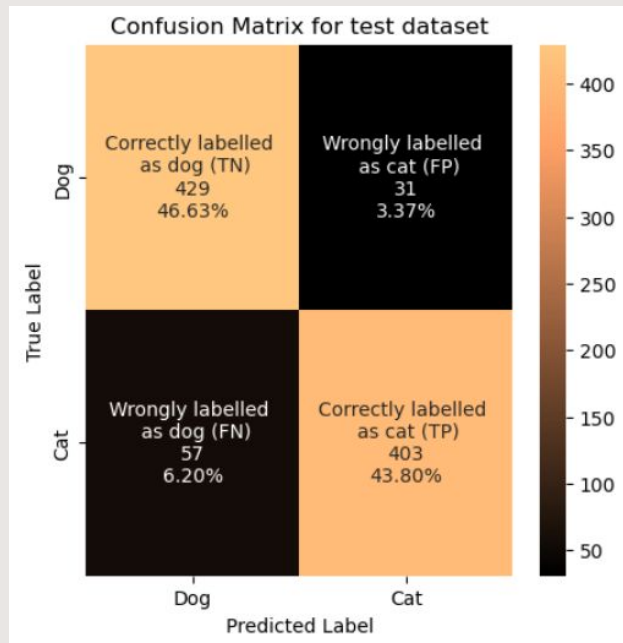
	Model	Data transformation to	Accuracy Score for train	Accuracy Score for test	F1 score
0	KNN (Baseline)	Tf-idf	0.865489	0.781522	0.789529
1	MultiNB	Tf-idf	0.955978	0.890217	0.891515
2	MultiNB	single_word	0.958152	0.902174	0.900222
3	MultiNB	2-word phrase	0.999728	0.840217	0.847668
4	BernNB	Tf-idf	0.963587	0.900000	0.897778
5	BernNB	single_word	0.963587	0.900000	0.897778
6	BernNB	2-word phrase	1.000000	0.825000	0.815578
7	GausNB	Tf-idf	0.971739	0.742391	0.751832
8	GausNB	single_word	0.964402	0.763043	0.777551
9	GausNB	2-word phrase	0.844565	0.773913	0.742574
10	Logistic Regression	Tf-idf	0.960054	0.885870	0.880546
11	Logistic Regression	single_word	0.999728	0.880435	0.878587
12	Logistic Regression	2-word phrase	1.000000	0.781522	0.759857
13	MultiNB with hyperparameter tuning	single_word	0.903261	0.904348	0.901566

*Selected model for tuning*

*Best result post-tuning*



# 90% of predictions are correctly labeled



Highly accurate model

- Total Observations: 920
- Correct classifications: 832

Low Misclassification

- Wrongly labeled as dog: 6.2%
- Wrongly labeled as cat: 3.3%

# Distinct words between cats and dogs

## Top 10 words for cats

Meowing

Meow

Swat

Litterbox

Galaxy

Jackson

Feliway

Diffuser

Tabby

Resident

Behaviours

Cat types



Tabby



## Top 10 words for dogs

Bark

Trainer

Flea

Diarrhea

Infection

Treatment

Collie

Poodle

Terrier

Daycare

Behaviours

Medical

Dog types

Poodle



7

# Conclusion and Recommendations

Our key takeaways



# Model Results Recap



**8000**

**Number of posts**

Extensive content that  
allow our model to be  
trained accurately

**90%**

**Accuracy**

Less than 10% of content  
is misclassified

**12%**

**Increase in accuracy**

New model is  
outperforming previous  
model by 12%

# Limitations

## Article availability

- Limited availability of articles at launch
  - Rapidly expand based on popular topics and what are people looking for

## Cat/Dog specific

- Cat and dogs are the most popular pets
  - Model currently limited to cat and dog classification

## Context

- Recommendations and articles are based on global trends
  - Not specific to Singapore

# Conclusion

Problem statement:

With an influx of **inexperienced** pet owners **overly reliant** on vets and pet store, how can we better optimize everyone's time?

## Refocus your business

- Focus on the core of your business
  - Treating pets
  - Driving sales growth

## Reduce dependance

- Provide users with extensive and comprehensive articles to cover all their basic needs

## Leverage insight

- Get information on customers/patients
- Based on their app usage

## Increase care

- Reduce abandonment rate
-

# Recommendations

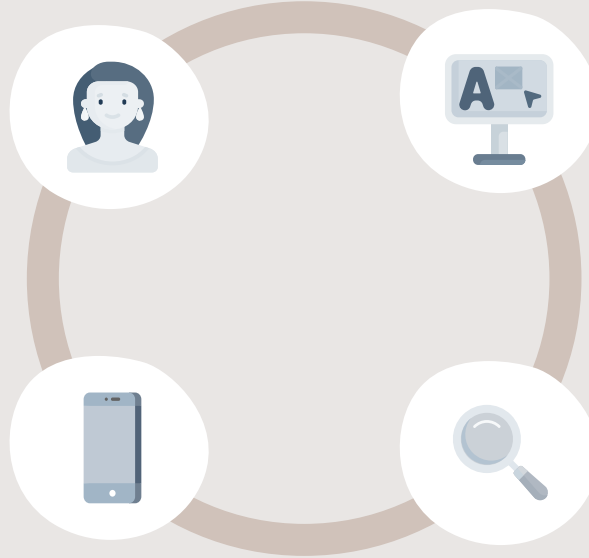


## Manpower

Adjust manpower needed to answer phone calls and emails and shift your attention to what really matters

## Digitalisation

Display QR codes or provide a device your customers can use to browse articles or get answers



## Communication

Improve on communications and website by focusing on popular and trending terms

## Inventory / Stocks

Keep up to date with current trends or outbreaks to always have stocks for critical and popular products

# Roadmap



**Q1 2023**

Offer weekly surveys to gather insights on users preferences and needs



**Q2 2023**

Provides insights from user usages.

- Popular question
- Survey results



**Q4 2023**

Organize events and workshops with end users and partners



**2025**

Deploy service to other countries

- Cater to local markets





**Thank you!**

Questions?

Reach us at  
**[petwhisperers@smartpet.com](mailto:petwhisperers@smartpet.com)**