Capstone project 1: Milestone Report

Project Title:
Pet Product Auto-Subcategorization by Review Analysis

< Abstract >

The goal of this project is to create a system that automatically classifies products in Pet Supplies category into subcategories by analyzing the reviews. In this project, the data collected on Amazon during 2014 - 2015 in the US is used. Data has been obtained, cleaned, and wrangled. Also, the exploratory analysis has been performed.

< Table of Contents >

1. Problem Statement

1.1. Problem

E-commerce companies set up categories for their products; for example, Clothing, Beauty, Books, Pet Supplies, etc. However, if the number of products in a category has been growing, they might want to classify the products into subcategories for several reasons. Because the number of products would be large, it could be difficult to categorize them one by one.

Nowadays, most e-commerce websites have reviews, which are written by customers and which help future customers decide whether they buy. On the other hand, reviews also provide much information to an e-commerce company; what customers liked or disliked, what they wanted, how they used it, and so on. In other words, reviews include information to categorize a product into subcategories.

Here, I chose Pet Supplies as a category reclassified into subcategories. I am going to make a system that automatically classifies products in Pet Supplies category into subcategories by analyzing the reviews.

1.2. Client

The first clients will be e-commerce companies that would like to use reviews to subcategorize their products. Their purposes could be to improve analysis of trends and customer needs to a specific field, and/or to increase customer satisfaction by easy access to a product they want.

2. Description of Dataset

2.1. Data Source

The animal product review data was acquired from AWS. This data contains information about the marketplace, product ID, product name, product category, star rating, review, date, etc. In this project, I focused on the data collected during 2014-2015 in the US.

- Amazon Pet Supplies reviews in the US (gz file) from AWS (1995 – 2015)
  https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Pet_Products_v1_00.tsv.gz

2.2. Data Cleaning

The following steps were performed:

- Extracting columns which were going to be used when loading the dataset,
- Extracting data collected in 2014 or 2015,
- Removing columns having a single value, and
- Dealing with missing values and duplicate data

There were 1,705,229 observations (reviews) and 128,995 products in the data at the beginning. Missing values were dropped because the rate was low. There were 216 (0.0013%) observations that lacked a review body. 18 products (0.0014%) were removed from the data by removing the observations (missing data). There were 8,004 duplicate rows. They were purely removed from the data. In consequence, 1,697,225 observations (reviews) remained, but the number of products in the data did not change (128, 977 products).

2.3. Data Wrangling

The following steps were performed:

- Removing short and long reviews
- Adjusting the number of reviews per product,
- Merging reviews to the same product in one
- Tokenization,
- Removing words other than alphabets,
- Removing stop words,
- Stemming and lemmatization,
- Removing words other than nouns,
- Removing words appearing in 5 or fewer products, and
- Removing products having 4 tokens or less

Observations that had a short review (fewer than 30 characters) and long reviews (more than 760 characters, 95th

percentile) were removed. In consequence, 12,177 (9.4%) products were dropped. A short review would not have enough information to subcategorize the product, and the products having only short reviews would not be very active. A long review would have too much information, and it makes the model complicated.

At this moment, I was not sure how many reviews per product would be enough to subcategorize a product. Of course, it depends on the length of a review and the number of characteristic words in a review. So, I decided to prepare three types of datasets depending on how many reviews a product had; 2 to 5 reviews per product (Dataset 1), 5 reviews (Dataset 2), and 10 reviews (Dataset 3). When a product had fewer reviews than the set amount, it was removed. When a product had more reviews than the set amount, the latest reviews were selected for the product. Then, the reviews to the same product were merged so that each product had one long review for the next step.

The next step was tokenization. The sentences were chopped into words. Then, words that would not be used to predict were handled. First, non-alphabet words and stop words were removed. Then, the other words were transformed into the base forms through stemming and lemmatization. Second, I determined that words other than nouns would not be necessary to achieve the goal, and removed non-noun words by using POS. Third, words appearing only in 5 times or fewer documents were dropped because the words would not only have much impact but also increase the number of dimensions and make the modeling complicated. Finally, the words in the Top 100 appearance that were not characteristic were removed by hand.

Also, products having fewer than 4 tokens after cleaning were removed. It was caused because some products had a few reviews and the reviews did not have enough characteristic information.

After this processing, the number of reviews and the products, and the number of unique tokens were following. Dataset 1 held 48% of the original number of products, but 25% (Dataset 2) and 15% (Dataset 3). When more products were expected to be subcategorized, Dataset 1 would be the best. On the other hand, Dataset 1 had fewer reviews per product, and I was not sure that the amount was enough. I will figure it out when they are applied to the model.

Table 1. Three types of datasets and the numbers

Dataset	Reviews / product	Total reviews	Total products	Total tokens (unique)
-	(Original*)	1,697,225	128,977 (100%)	-
1	2 to 5	245,565	61,770 (48%)	8,032
2	5	161,765	32,295 (25%)	6,385
3	10	195,570	19,557 (15%)	6,840

* After removing missing values and duplicate data

Just in case, all products that were removed from the data were shown as tables in the Jupyter Notebook.

3. Findings from the exploratory analysis

3.1. Tokens

There were 8,032 (Dataset 1), 6,385 (Dataset 2), and 6,840 (Dataset 3) unique words in each Dataset after preprocessing. Words (tokens) about animal kinds and application types appeared high up on the list. For example, animal kinds: dog, cat, and bird. Application kinds: food, toy, treat, tank, collar, cage, leash, bowl, etc. These words will be a significant clue to subcategorize products. I picked up Dataset 1 here (Figure 1), but the other two datasets were similar.

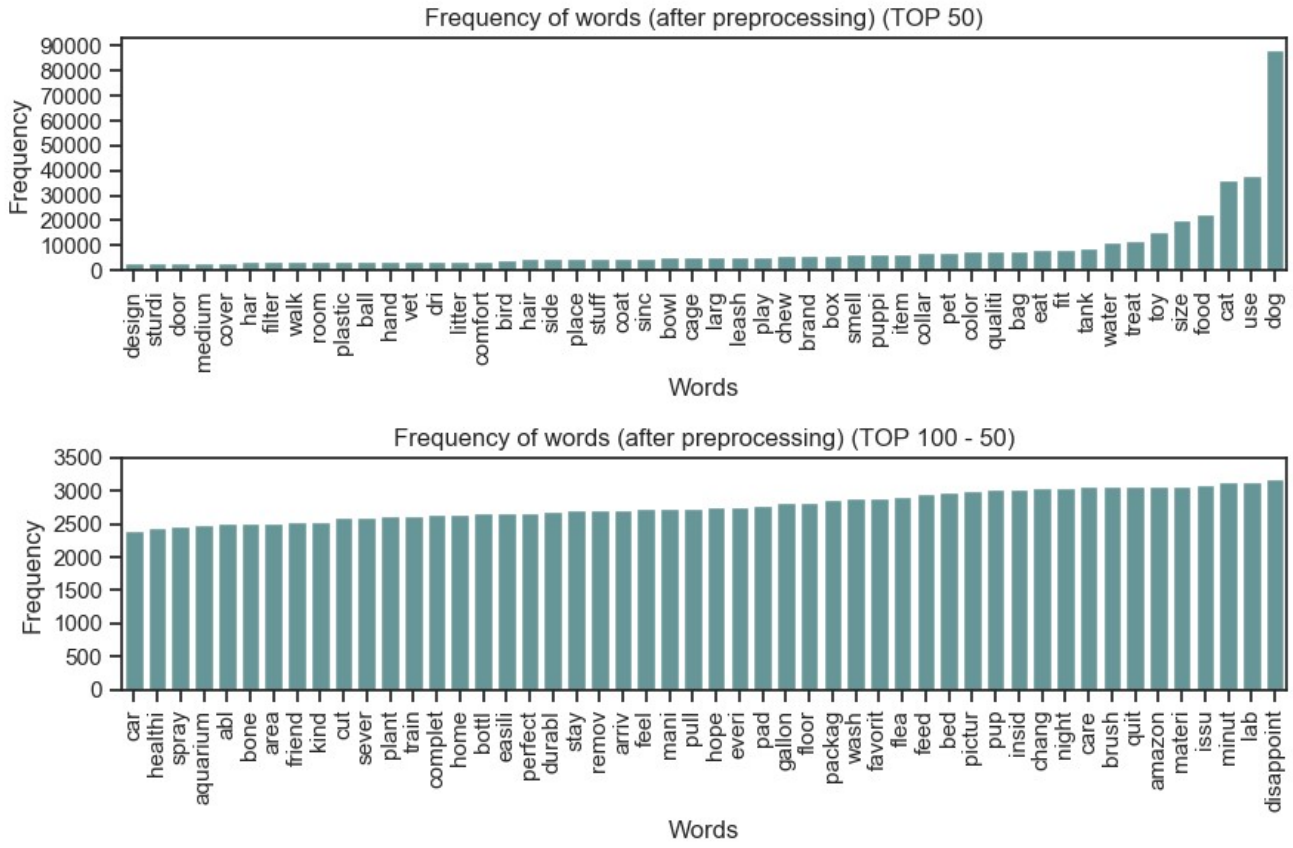

Figure 1. Frequency of words (Top 100, Dataset 1)

Next, the number of tokens per product were counted. The results were shown in Table 2 and Figure 2.

Table 2. The summary statistics of the number of tokens per product

Dataset	Min	25%	50%	75%	90%	Max	Mean	SD
1	5	14	23	35	47	119	26.2	15.8
2	5	20	28	39	50	130	30.5	14.6
3	9	46	59	76	94	234	62.5	24.0

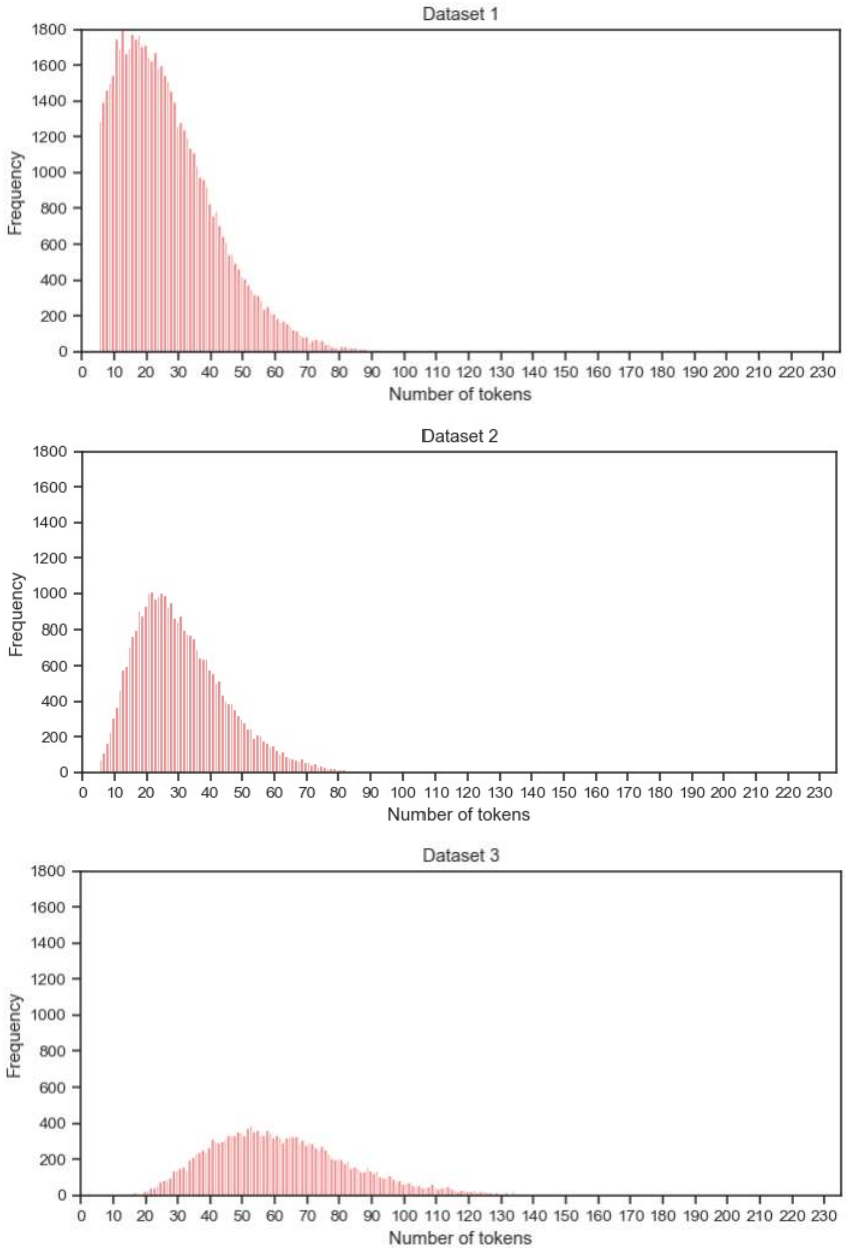

Figure 2. The number of tokens per product

There were 61,770 (Dataset 1), 32,295 (Dataset 2), and 19,557 (Dataset 3) products in each Dataset after preprocessing. Because each Dataset had a different number of products, the heights of the bar plots were different (Figure 4).

In Dataset 1, 90% of the products had 47 or fewer tokens, and the mean was 26.2. Datasets 2 and 3 had the larger number of both the 90th percentile and the mean. This was because Dataset 2 or Dataset 3 had 5 or 10 reviews per product, respectively, but 2 to 5 reviews in Dataset 1. The mean of tokens of each product in Dataset 3 was about twice than in Dataset 1 or 2 (Table 2).

3.2. The number of products relating to each animal kind

Several animals were found in the tokens above. Here, I counted the rough numbers of products relating to each specific animal. I picked up some animals: dogs, cats, birds, fish, and rabbits. When a product had a token such as ‘dog’, the product was counted as a product relating to dogs. The keywords I chose for each category were shown in Table 3.

Table 3. The keywords for each animal category

Category	Keywords
Dog	dog, puppi, doggi
Cat	cat, kitti, kitten
Bird	bird, chick
Fish	fish
Rabbit	rabbit, bunni

One product could be counted as both dogs related and cats related when the product had both tokens; ‘dog’ and ‘cat’. It might mean the product could be used for dogs and cats. Figure 3 showed the number of products relating to each animal in Dataset 1.

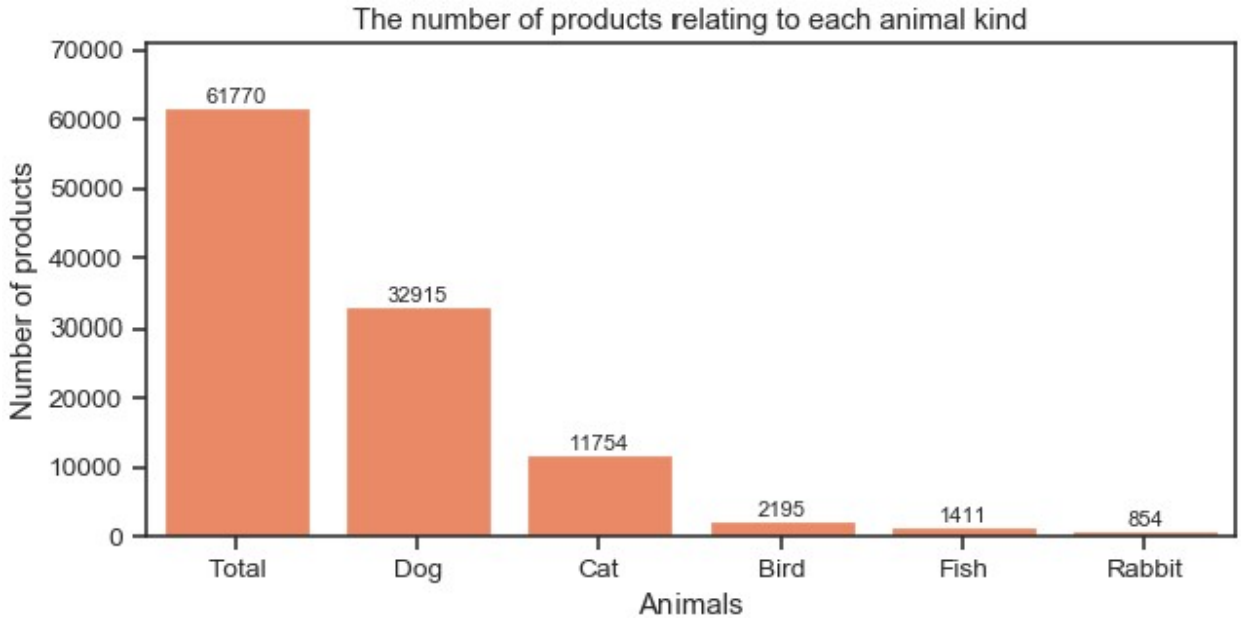

Figure 3. The number of products relating to each animal kind (Dataset 1)

Dog and cat products were the largest two groups. Especially, 53% of the products were related to dogs. The trends in Dataset 2 and 3 were the same as Dataset 1.

3.3. The number of products relating to each application type

There were tokens such as food, toy, treat, tank, collar, cage, leash, bowl, etc. in the data. Here, I prepared 7 categories for application types. The keywords for each category were shown in Table 4.

Table 4. The keywords for each application category

Categories	Keywords
Toy	toy, tunnel, ball, rope, stuff
Food	food, dri, wet
Treat	treat, snack, cooki
Collar & Leash	collar, leash
Clothes	shirt, coat, sweater, costum
Cage	cage, crate, carrier, kennel
Toilet	litter, pad

Each category was counted at the same rule as counting animal kinds. The result was in Figure 4.

The Top 2 categories were Toy and Food. However, the products were relatively spread in the seven categories. Also, I picked up the categories that would mainly fit products for dogs and cats. There would be unique products that were not classified in them, especially products for fish or other animals. The trends in Dataset 2 and 3 were the same as Dataset 1.

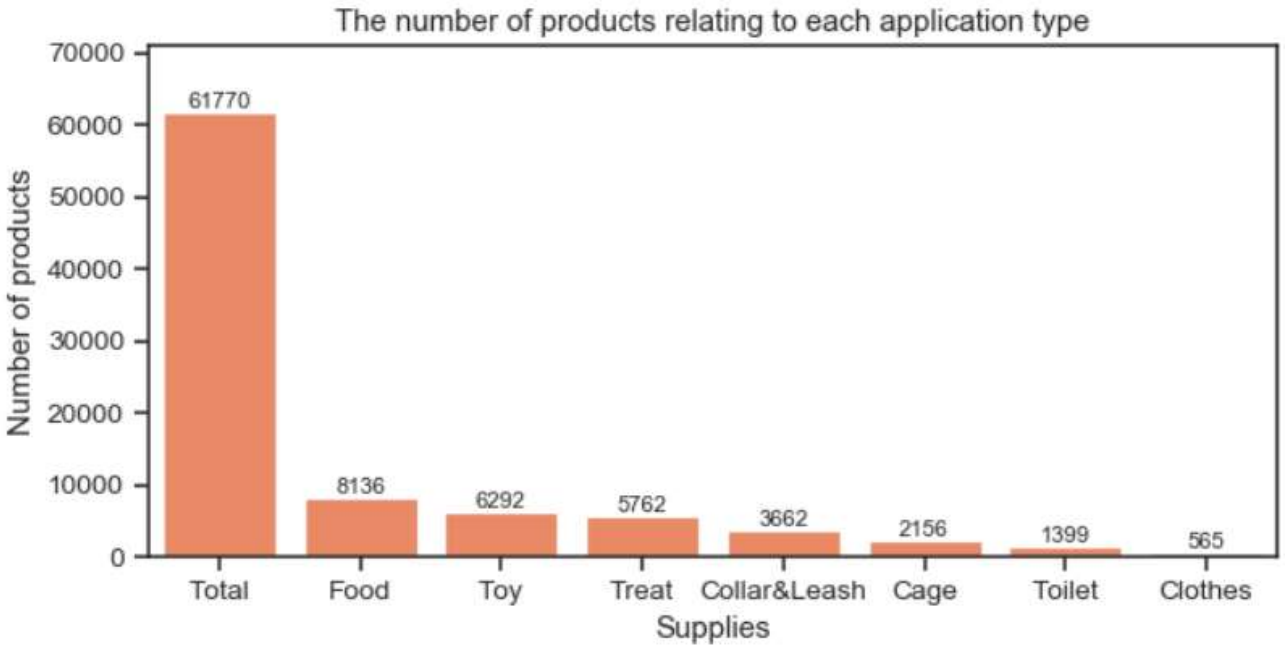

Figure 4. The number of products relating to each application category (Dataset 1)

3.4. The rough relationship visualization between an animal kind and application type

I picked up Dog and Cat as animal categories, and Toy and Food as application categories. Then, the relationships between them were checked. This meant that a product was classified as a dog- and toy-related product if the product had tokens such as ‘dog’ and ‘toy’. The result was shown in Figure 5.

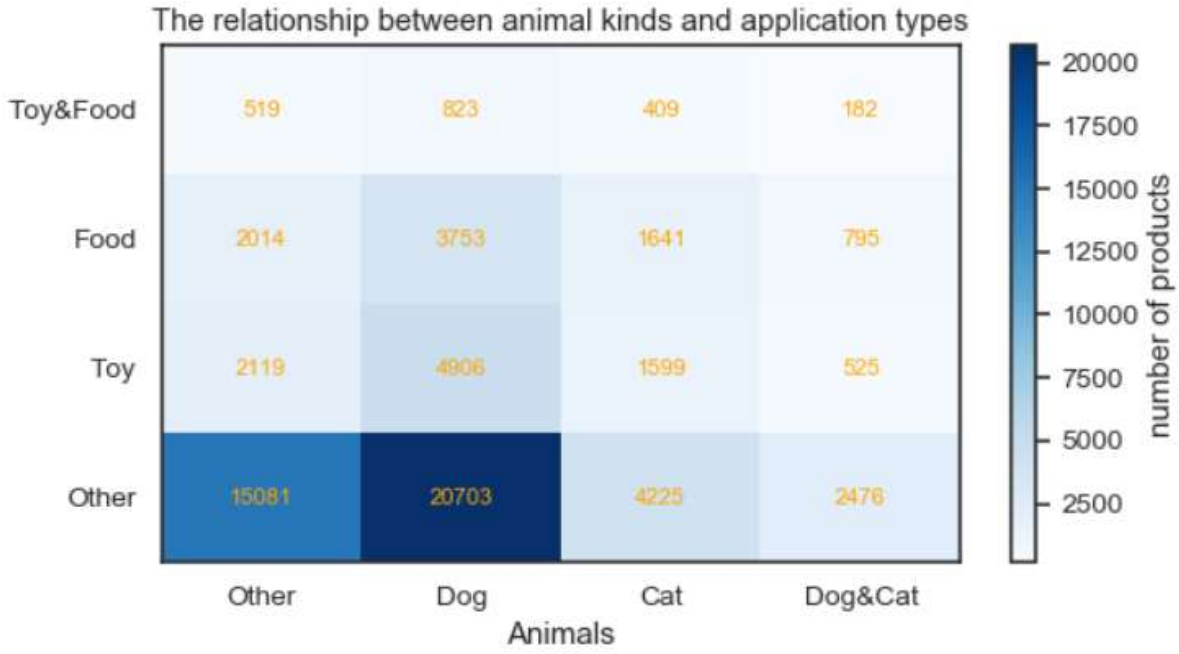

Figure 5. The relationship between animal kinds and application types (Dataset 1)

Food and Toy for dogs were the largest two categories on the chart, but many products were in Other for dogs or Other for the other animals. I will categorize them more precisely in the modeling part of this project. The trends in Dataset 2 and 3 were quite similar to Dataset 1.