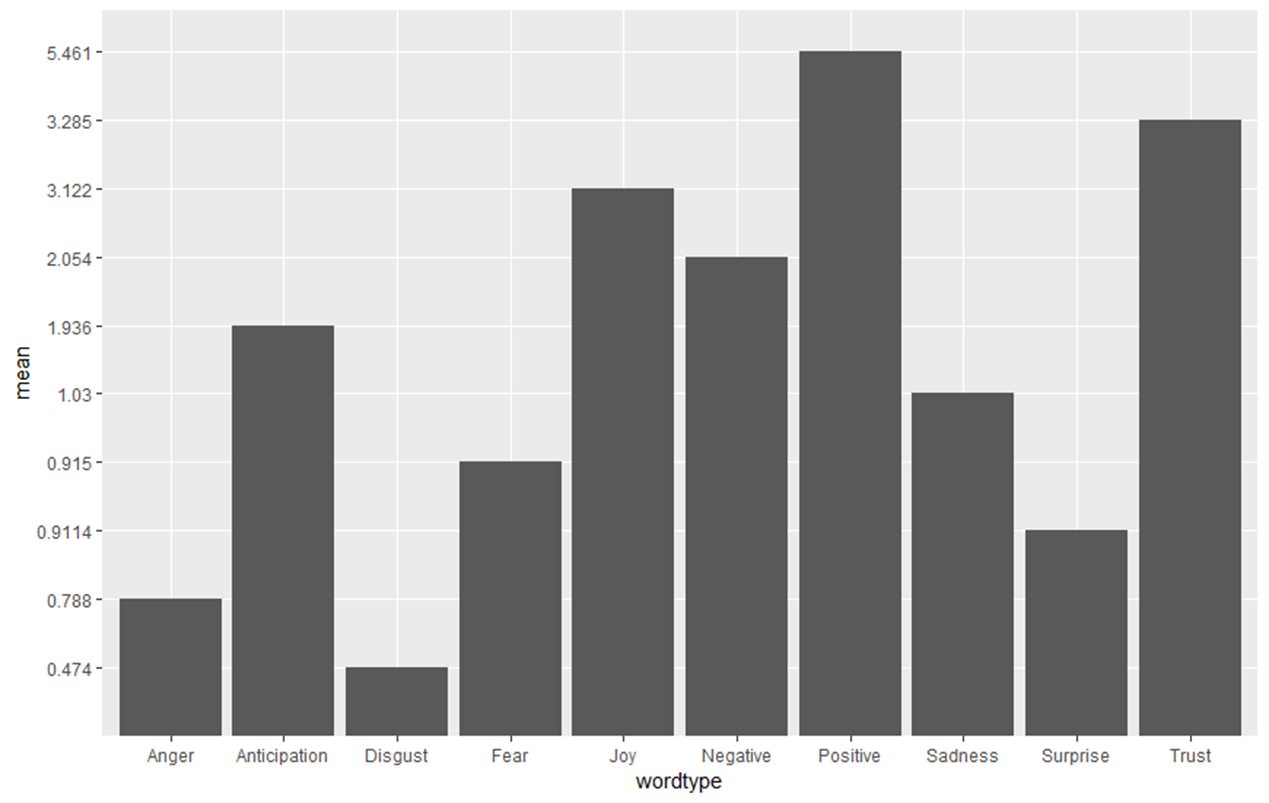
Kaggle.com is currently running a contest to predict the time to adoption for pets listed on the PetFinder.MY website. The data provided is photographic, semi-ordered and features such as the age of the animal and a written description. The photographic data needed to be processed so machine learning algorithms can analyze the data. While sentiment data was provided it was decided that the descriptions would be used and text analysis done using a lexicon would be used. Three machine learning algorithms were used. Random Forest, Naïve Bayes and K-Nearest Neighbors. Of these Random Forest done on the non-photographic data only provided the best results. The photographic data was problematic due to the inconsistency of the quality of the photographs.

The photographs were converted to a vector of values using the Single Value Decomposition method. This converts the RGB data into processable data points. The most popular way to classify this kind of photographic data is using neural networks. Since the analysis would be using methods besides neural networks a test on fruit classification was run using Random Forests. RF at least is able to properly identify pictures of fruit and so this analysis went ahead with other machine learning algorithms.

To manage the vector size the data was reduced to 50 x 50 pixels, it is hoped that enough detail is preserved in these photos that they are useful. This also spares computation time. Red Blue and Green data are preserved, it is thought color will be a significant factor in how the photographs are received. This results in vectors 7500 data points long. Photos that could not be converted were not used and that entry ignored.

When looking over the Google generated data it was found that the sentiment analysis provided was either a positive or negative entry. It was decided that a more detailed analysis would be beneficial. A lexicon was found which included words from several languages and included sentiments like trust, sadness, anger as well as positive and negative sentiment. The NRC Word-Emotion Association Lexicon was used since this includes sentiments such as trust, joy and sadness along with the positive and negative emotions. The total number of times a sentiment appears is added up to create the data. For example if words that evoke sadness appear twice then a sadness score of 2 appears. The graph below shows the average number of appearances of the sentiments in each description.



Processing the descriptions for sentiment analysis was done by breaking the descriptions down into individual words, removing stop words such as, and, or, that, and the. Then removing any words that did not appear in the lexicon.

The word cloud shows graphically the appearance of all words in the lexicon.

A close up of text on a white background

Description automatically generated

The rest of the data was analyzed for any clear signs of relationships. Cats were more likely to be adopted in the first few weeks than dogs were. Yet in the importance table from the Random Forest analysis this did not figure prominently. Part of the possible reason is size. Yet Cats make up almost as many small and medium classifications as dogs. Medium and Large are almost identical. One would expect that Maturity Size and Type would have a significant effect on which pet gets adopted. This suggests that interaction variables may enhance accuracy. Such as pet is cat and small, dog and large and so on. Small and medium sized animals did vary somewhat.

Three models were successfully used in creating forecasts. Naïve Bayes, K Nearest Neighbors and Random Forest. The measure we will use to check the usefulness of the forecast is the Cohen’s Kappa measure. This measure tries to take into account the chances of getting the right answer randomly. We use the guidelines from Statistical Methods for Rates and Proportions which ranks 0 to .40 as a poor match, .40 to .75 as a fair match and 75 and over as an excellent match. This was chosen to prevent overstating the significance of the match.

None of the used machine learning algorithms performed well. Random forest worked best and was able to accurately predict the data more than 50% of the time. Naïve Bayes was the worst being no better than a random guess. KNN was able to accurately forecast but received a poor rating from our rating system.

Photographs vary in quality and in type as shown earlier. It was decided that the photographic data should be classified. K-Means was used to classify the data. We see that the addition of the photographic data degraded for two out of three methods. For the Naïve Bayes modle the improvement was not enough to provide a fair matching qualification. In this analysis the photographic data was essentially useless. This was likely due to the need for further processing and classification for the photographic data. The results of the Naïve Bayes hints that this may be worth the additional work required to classify the data to get better results.

Processing the data proved to be the greatest challenge in this project. During this project it was learned image analysis is an entire sub-filed in data science and photographic data must be processed and classified correctly in order to get good results. Further work on this project could include improved photographic processing. Text processing could also potentially be improved with using n-grams greater than 1. If neural networks could be applied properly then it’s possible this will beat random forest. Reason being is while random forest simulates a decision-making process to a certain extent neural networks are the preferred method for this kind of analysis particularly where photographs are concerned.