Data

The data used for this analysis is a series happy moments. In collaboration with the University of Tokyo, MIT researchers interviewed thousands of people and asked them to list 10 happy moments that occurred within the last 24 hours. Their responses were recorded and compiled into HappyDB, a corpus of 100,535 happy moments. The data set is host publicly on github, as a zip file (https://rit-public.github.io/HappyDB/).

It is important to note that this curated data set contains cleaned textual data. In it includes 9 variables, many of which identify the author and qualities of the texts. The variables of interest in this analysis are cleaned_hm, the list of happy moments, and its associated predicted category. For my analysis, I sampled 15,000 of these responses and trained on 80% of the data.

Response	Predicted Category
I was happy when my son got 90% marks in his examination	Acheivement
I went to the gym this morning and did yoga	Excercise
My dad and I went fishing	Bonding

Data Preprocessing:

In order to analyze how words influence a predicted category, the data set had to be transformed. First, each character vector response is converted into a corpus. A corpus is simply a collection of natural langauge constructed with a specific purpose. Our corpus consists of 15,000 responses samples from the larger dataset.

```
as.character(corpus[[1]])
```

```
## [1] "i went to the farm store and found fruit trees on sale for 70% off. "
```

Entries in the corpus are then cleaned of symbols, punctuations, and stop words. Once accomplished the corpus is reduced a collection of key words.

```
as.character(a[[1]])
```

```
## [1] "    farm store   fruit trees  sale   "
```

The next step is to create a document term matrix. A document term matrix is a corpus transformation that represents each word as a feature, and each response as a row. The matrix is populated with values 0 or 1, depending on the absence or presence of that word in the document. Traditionally, document term matrices are high dimensional due to thousands of words that can be used as a feature. Often times many of these cells are empty to represent the absence of a word from that document. High sparsity (the proportion of entries which are zero of the tdm) is another known condition of document terms matrices.

```
##     Terms
## Docs farm fruit sale store
##    1    1     1    1     1
##    2    0     0    0     0
##    3    0     0    0     0
```

Exploratory Data Analysis:

In this section, I present some preliminary results from summaries and exploratory analysis of the cleaned data obtained above. First we look at the response variable **predicted category**. Next, we explore statistics regarding the explanatory variables **terms**.

Predicted Category

Affection and achievement were the most talked about categories (33.9% and 34.1% respectively). This follows conventional logic given that love and sense of accomplishment are vital traits to a good quality of life. Nature and exercises accounted for just 1.2% and 1.8% of labeled responses in the entire data set.

In a predictive setting, having highly disproportionate classes can lead to errors when making predictions based on that training data. When classes are unequally represented, models may simply decide to always predict a certain class in order to achieve a high accuracy. This accuracy paradox does not reflect effective predictions for the model, but rather the state of the underlying distribution. While this is not the case for our data set, consideration must be taken for the classifications of minority labels.

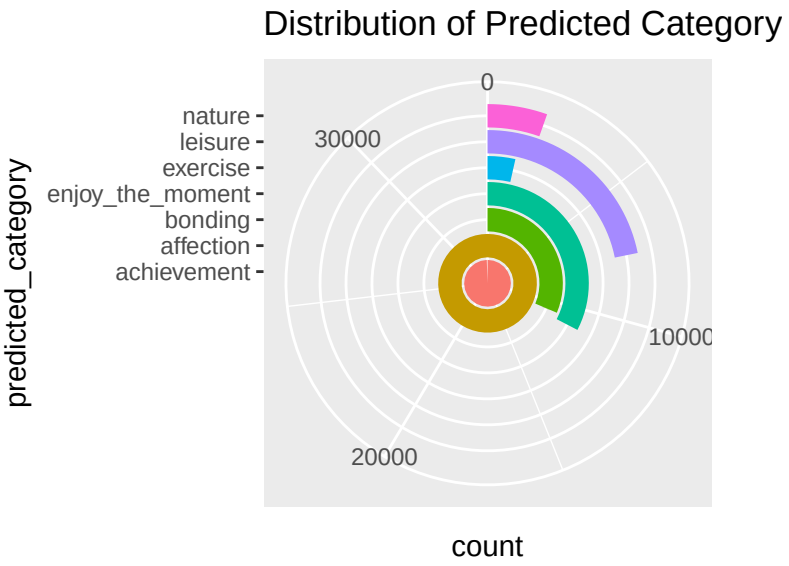

Figure 1: PieChart

Term Count

As shown in the figure below, the majority of entries are between 5 and 14 words. Over 4,000 reviews contain over 50 words. Many of these words add little to no value to an algorithmic interpretation of the sentence. These are the commonly used words such as “a”, “I”, “was” that were removed from the corpus.

Terms used in each category

Two metrics frequently used to quantify the importance of a word are it’s term frequency (tf), and inverse document frequency (idf). Tf measures how frequently a word appears in a document, while idf weights frequently used words less than words rarely used (Leopold, 2002). When combined, the tf-idf is the frequency of a term adjusted by how rarely it is used.

```
## Selecting by tf_idf
```

As shown in the figure above, a breakdown of the highest tf-idf word per category yields further insight into words associated with each label. For phrases related to achievement, words like “won”, “received”, and “accepted” have the highest adjusted usage. In responses categorized as exercise, we see terms related to working out and equipment used. These key terms provide natural separators by which our algorithms can uses to classify future responses. In the next section, we discuss some of the methods used to predict the categories of happy moments.

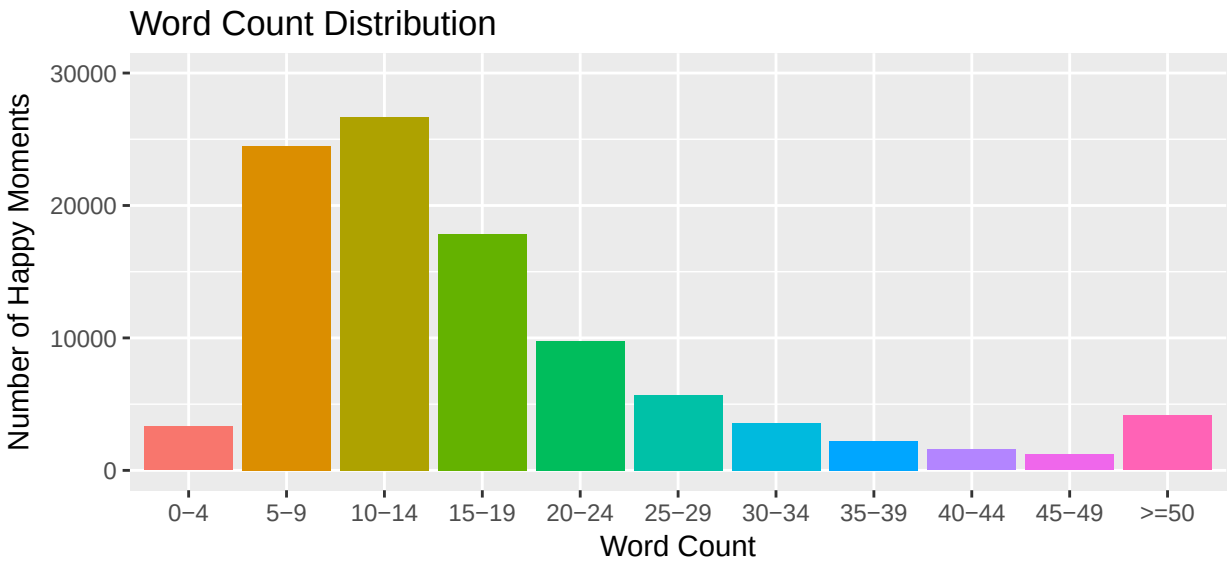

Figure 2: Word Count Distribution

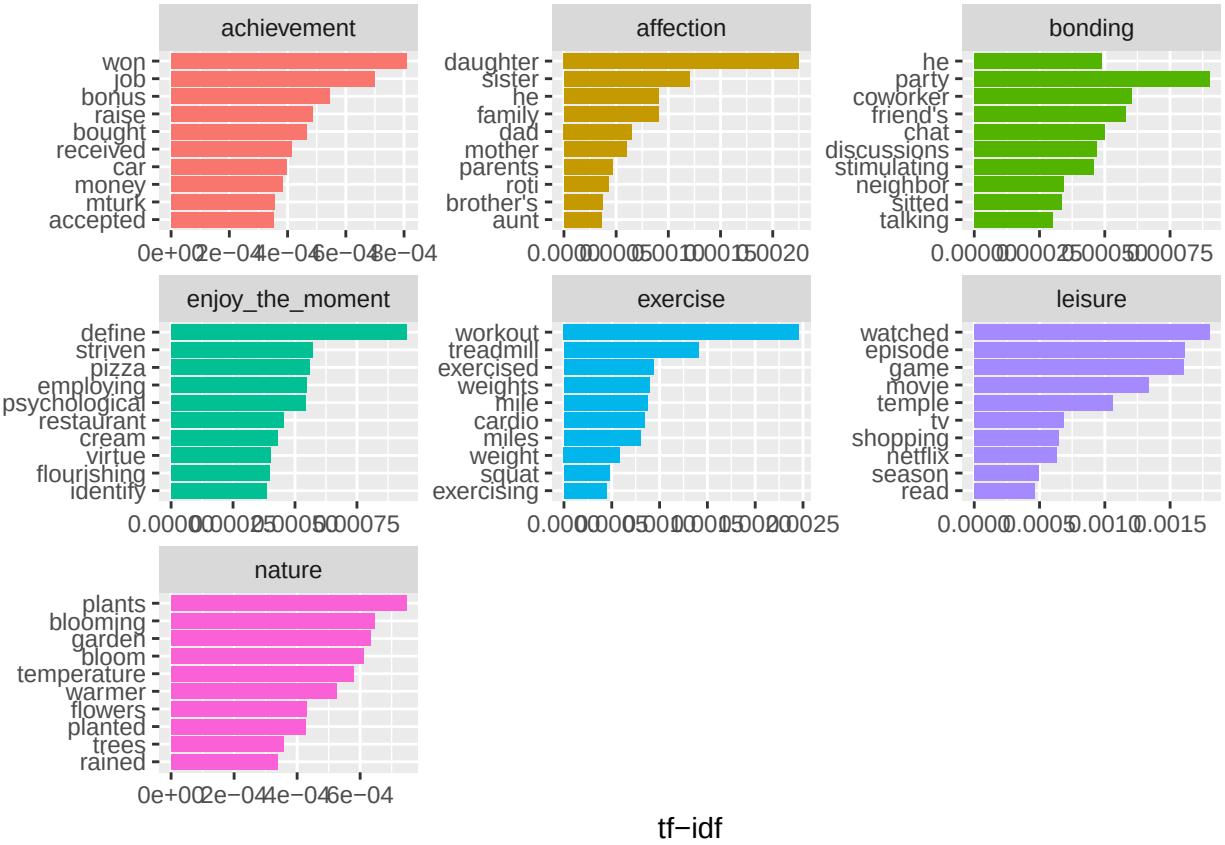

Figure 3: TD-IDF per Category

Figure 4: Framework of Text Classification