**A. Introduction**

The growth of Crossfit has been one of the biggest developments in the fitness industry over the past decade. Promoted as both a physical exercise philosophy and also as a competitive fitness sport, Crossfit is a high-intensity fitness program incorporating elements from several sports and exercise protocols such as high-intensity interval training, Olympic weightlifting, plyometrics, powerlifting, gymnastics, strongman, and so forth. Now with over 10,000 Crossfit affiliated gyms (boxes) throughout the United States, the market has certainly become more saturated and gyms must initiate more unique marketing strategies to attract new members. In this post, I will investigate how some prominent Crossfit boxes are utilizing Twitter to engage with consumers. While Twitter is a great platform for news and entertainment, it is usually not the place for customer acquisition given the lack of targeted messaging. Furthermore, unlike platforms like Instagram,Twitter is simply not an image/video centric tool where followers can view accomplishments from their favorite fitness heroes, witness people achieving their goals, and so forth. Given these shortcomings, I wanted to understand how some prominent Crossfit boxes are actually using their Twitter accounts. In this post, I will investigate tweeting patterns, content characteristics, perform sentiment analysis, and build a predictive model to predict retweets.

**B. Extract Data From Twitter**

We begin by extracting the desired data from Twitter using the rtweet package in R. There are six prominent Crossfit gyms whose entire Twitter timeline we will use. To get this data, I looped through a vector containing each of their Twitter handles and used the get\_timeline function to pull the desired data. Notice that there is a user defined function called add\_engineered\_features that is used to add a number of extra date columns. That function is available on my GitHub page [here](https://gist.github.com/abmathewks/48141aae0fc77923bb051137c689a173).

library(rtweet)

library(lubridate)

library(devtools)

library(data.table)

library(ggplot2)

library(hms)

library(scales)

# set working directory

setwd("~/Desktop/rtweet\_crossfit")

final\_dat.tmp <- list()

cf\_gyms <- c("reebokcrossfit5", "crossfitmayhem", "crossfitsanitas", "sfcrossfit", "cfbelltown", "WindyCityCF")

for(each\_box in cf\_gyms){

message("Getting data for: ", each\_box)

each\_cf % data.table()

each\_cf$crossfit\_name <- each\_box

suppressWarnings( add\_engineered\_dates(each\_cf, date\_col = "created\_at") )

final\_dat.tmp[[each\_box]] <- each\_cf

message("")

}

final\_dat <- rbindlist(final\_dat.tmp)

final\_dat$contains\_hashtags <- ifelse(!is.na(final\_dat$hashtags), 1, 0)

final\_dat$hashtags\_count <- lapply(final\_dat$hashtags, function(x) length(x[!is.na(x)]))

**C. Exploratory Data Analysis**

Let us start by investigating this data set to get a better understanding of trends and patterns across these various Crossfit boxes. The important thing to note is that not all these twitter accounts are currently active. We can see that crossfitmayhem, sfcrossfit, and WindyCityCF are the only ones who remain active.

final\_dat[, .(min\_max = range(as.Date(created\_at))), by=crossfit\_name][,label := rep(c("first\_tweet","last\_tweet"))][]

***C1. Total Number of Tweets***

Sfcrossfit, which is the oldest of these gyms, has the highest number of tweets. However, when looking at the total number of tweets per active month, they were less active than two other gyms.

## total number of tweets

p1 = final\_dat[, .N, by=crossfit\_name] %>%

ggplot(., aes(x=reorder(crossfit\_name, N), y=N)) + geom\_bar(stat='identity', fill="steelblue") +

coord\_flip() + labs(x="", y="") + ylim(0,3000) + ggtitle("Total Number of Tweets") +

theme(plot.title = element\_text(hjust = 0.5),

axis.ticks.x = element\_line(colour = "black"),

axis.ticks.y = element\_line(colour = "black"))

## number of tweets per active month

p2 = final\_dat[, .(.N, start=lubridate::ymd\_hms(min(created\_at)), months\_active=lubridate::interval(lubridate::ymd\_hms(min(created\_at)), Sys.Date()) %/% months(1)), by=crossfit\_name][,

.(tweets\_per\_month = N/months\_active), by=crossfit\_name] %>%

ggplot(., aes(x=reorder(crossfit\_name, tweets\_per\_month), y=tweets\_per\_month)) +

geom\_bar(stat='identity', fill="steelblue") + coord\_flip() + labs(x="", y="") + ylim(0,32) +

theme(plot.title = element\_text(hjust = 0.5),

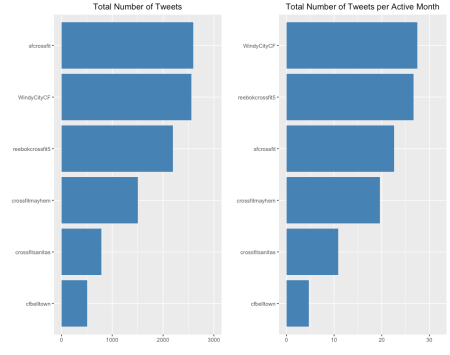
axis.ticks.x = element\_line(colour = "black"),

axis.ticks.y = element\_line(colour = "black")) +

ggtitle("Total Number of Tweets per Active Month")

## add both plots to a single pane

grid.arrange(p1, p2, nrow=1)

[](https://mathewanalytics.files.wordpress.com/2018/12/Rplot.png)

***C2. Total Number of Tweets Over Time***

The time series for the total number of tweets by month shows that each gym had one or two peaks from 2012 through 2016 where they were aggressively sharing content with their followers. However, over the past two years, each gym has reduced their twitter usage significantly.

## total number of tweets by month

final\_dat[, .N, by = .(crossfit\_name, created\_at\_YearMonth)][order(crossfit\_name, created\_at\_YearMonth)][,

created\_at\_YearMonth := lubridate::ymd(paste(created\_at\_YearMonth, "-01"))] %>%

ggplot(., aes(created\_at\_YearMonth, N, colour=crossfit\_name)) + geom\_line(group=1, lwd=0.6) +

facet\_wrap(~crossfit\_name) + labs(x="", y="") + theme(legend.position="none") +

theme(plot.title = element\_text(hjust = 0.5),

axis.ticks.x = element\_line(colour = "black"),

axis.ticks.y = element\_line(colour = "black"),

strip.text.x = element\_text(size = 10)) +

ggtitle("Total Number of Tweets")

## total number of tweets by year

ggplot(data = final\_dat,

aes(lubridate::month(created\_at, label=TRUE, abbr=TRUE),

group=factor(lubridate::year(created\_at)), color=factor(lubridate::year(created\_at))))+

geom\_line(stat="count") + geom\_point(stat="count") +

facet\_wrap(~crossfit\_name) + labs(x="", colour="Year") + xlab("") + ylab("") +

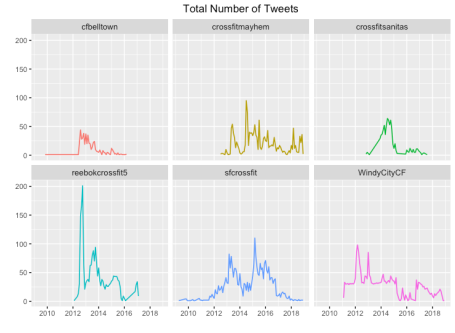
theme(plot.title = element\_text(hjust = 0.5),

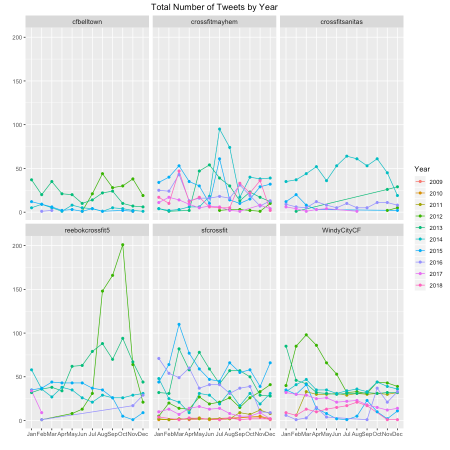
axis.ticks.x = element\_line(colour = "black"),

axis.ticks.y = element\_line(colour = "black"),

strip.text.x = element\_text(size = 10)) +

ggtitle("Total Number of Tweets by Year")

[](https://mathewanalytics.files.wordpress.com/2018/12/EDA_3-1.png)

[](https://mathewanalytics.files.wordpress.com/2018/12/EDA_3b.png)

***C3. Tweeting Volume by Year, Month, and Day***

For each Crossfit gym, I plotted the volume of tweets by year, month, and day. Oddly enough, there really are not any noticeable patterns in these charts.

## years with the highest number of tweets

ggplot(final\_dat, aes(created\_at\_Year)) + geom\_bar(fill="steelblue") +

facet\_wrap(~crossfit\_name) + labs(x="", y="") +

theme(plot.title = element\_text(hjust = 0.5),

axis.ticks.x = element\_line(colour = "black"),

axis.ticks.y = element\_line(colour = "black"),

strip.text.x = element\_text(size = 10)) + ylim(0,800) +

ggtitle("Total Number of Tweets by Year")

## months with the highest number of tweets

final\_dat[, created\_at\_YearMonth2 := lubridate::ymd(paste(created\_at\_YearMonth, "-01"))][] %>%

ggplot(., aes(lubridate::month(created\_at\_YearMonth2, label=TRUE, abbr=TRUE))) + geom\_bar(fill="steelblue") +

facet\_wrap(~crossfit\_name) + labs(x="", y="") +

theme(plot.title = element\_text(hjust = 0.5),

axis.ticks.x = element\_line(colour = "black"),

axis.ticks.y = element\_line(colour = "black"),

strip.text.x = element\_text(size = 10)) + ylim(0,500) +

ggtitle("Total Number of Tweets by Month")

## days with the highest number of tweets

final\_dat[, created\_at\_YearMonth2 := lubridate::wday(lubridate::ymd(paste(created\_at\_YearMonth, "-01")), label=T)][] %>%

ggplot(., aes(created\_at\_YearMonth2)) + geom\_bar(fill="steelblue") +

facet\_wrap(~crossfit\_name) + labs(x="", y="") +

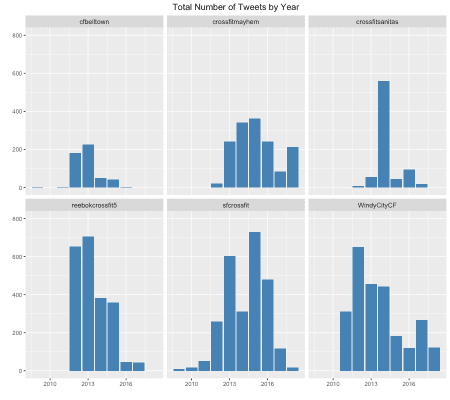
theme(plot.title = element\_text(hjust = 0.5),

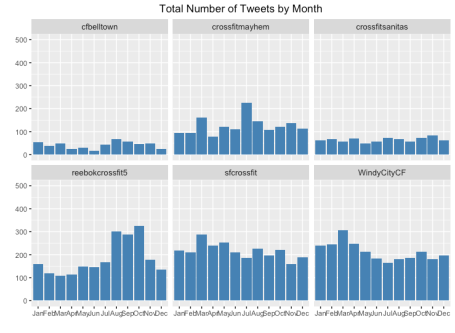
axis.ticks.x = element\_line(colour = "black"),

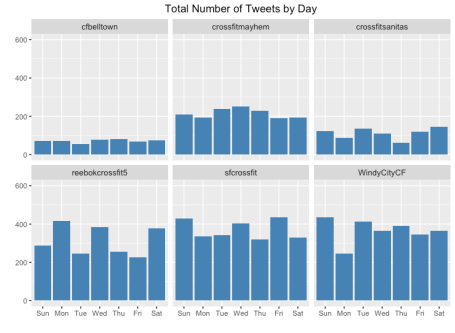
axis.ticks.y = element\_line(colour = "black"),

strip.text.x = element\_text(size = 10)) + ylim(0,600) +

ggtitle("Total Number of Tweets by Day")

[](https://mathewanalytics.files.wordpress.com/2018/12/EDA_10.png)

[](https://mathewanalytics.files.wordpress.com/2018/12/EDA_4.png)

[](https://mathewanalytics.files.wordpress.com/2018/12/EDA_5.png)

***C4. Tweeting Volume by Time of Day***

For most of these gyms, we see that the majority of their tweets came during the second half of the day or early in the morning. Given that most Crossfit boxes have their first classes at around 5am or 6am, and are usually most busy in the evening, this indicates that these businesses are tweeting during those hours where their facility is busiest.

## tweet density over the day

ggplot(data = final\_dat) +

geom\_density(aes(x = created\_at\_Time, y = ..scaled..),

fill="steelblue", alpha=0.3) +

xlab("Time") + ylab("Density") +

scale\_x\_datetime(breaks = date\_breaks("2 hours"),

labels = date\_format("%H:%M")) +

facet\_wrap(~crossfit\_name) + labs(x="", y="") +

theme(plot.title = element\_text(hjust = 0.5),

axis.text.x=element\_text(angle=90,hjust=1)) +

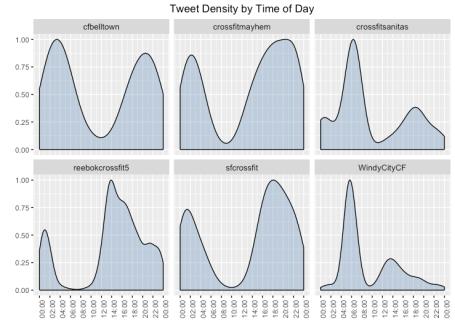
theme(plot.title = element\_text(hjust = 0.5),

axis.ticks.x = element\_line(colour = "black"),

axis.ticks.y = element\_line(colour = "black"),

strip.text.x = element\_text(size = 10)) +

ggtitle("Tweet Density by Time of Day")

[](https://mathewanalytics.files.wordpress.com/2018/12/EDA_6.png)

***C5. Source of Tweets***

While a couple of the gyms are using a marketing management platform like Hootsuite and IFTTT, the majority of Tweets are coming from a Twitter application or some other social media tool such as Facebook and Instagram.

## source of tweets

final\_dat[, .N, by = .(crossfit\_name, source)][order(crossfit\_name,-N)][,

head(.SD, 3),by=crossfit\_name] %>%

ggplot(., aes(x=source, y=N)) +

geom\_bar(stat="identity", fill="steelblue") + coord\_flip() +

facet\_wrap(~crossfit\_name) + labs(x="", y="") +

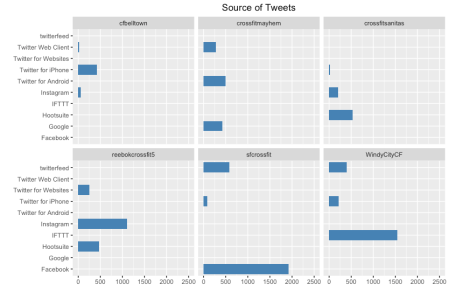
labs(x="", y="", colour="") + ylim(0,2500) +

theme(plot.title = element\_text(hjust = 0.5),

axis.ticks.x = element\_line(colour = "black"),

axis.ticks.y = element\_line(colour = "black")) +

ggtitle("Source of Tweets")

[](https://mathewanalytics.files.wordpress.com/2018/12/eda_7.png)

**D. Content Characteristics**

Let us investigate the characteristics of the content that is being tweeted. From looking at average tweet length to the amount of retweets, this information will further expand our understanding of how these prominent Crossfit boxes are utilizing their Twitter accounts.

***D1. Frequency of Retweets and Original Tweets***

Only one gym, cfbelltown, had a majority of their tweets being retweets. Furthermore, a plurality of tweets from crossfitmayhem were not original. In the plots showing retweets over time, we can also see that crossfitmayhem did a lot of retweeting from 2014 through 2016, but started pushing original tweets from then on out.

## retweets and original tweets

final\_dat[, .N, by=.(crossfit\_name,is\_retweet)][, percen := N/sum(N), by=.(crossfit\_name)][is\_retweet=="TRUE"] %>%

ggplot(., aes(x=reorder(crossfit\_name, N), y=percen)) +

geom\_bar(stat="identity", position ="dodge", fill="steelblue") + coord\_flip() +

labs(x="", y="", colour="") + ylim(0,0.8) +

theme(plot.title = element\_text(hjust = 0.5),

axis.ticks.x = element\_line(colour = "black"),

axis.ticks.y = element\_line(colour = "black")) +

ggtitle("Percentage of Tweets That Were Not Original")

# total number of retweets over time

ggplot(data = final\_dat, aes(x = created\_at, fill = is\_retweet)) +

geom\_histogram(bins=48) +

labs(x="", y="", colour="") +

scale\_fill\_manual(values = c("steelblue2","steelblue"), name = "Retweet") +

facet\_wrap(~crossfit\_name) + labs(x="", y="") +

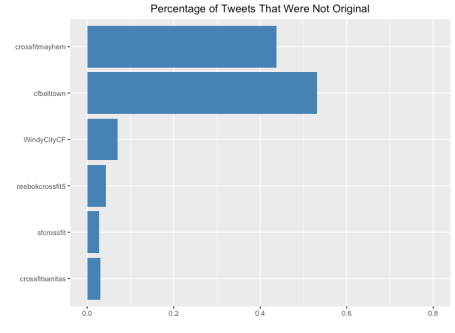
theme(plot.title = element\_text(hjust = 0.5),

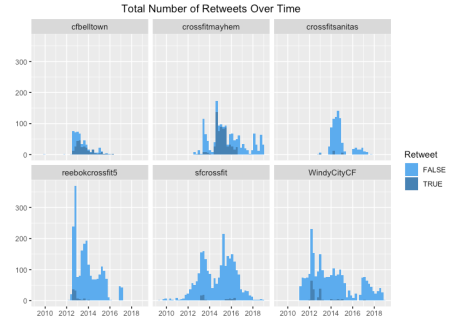
axis.ticks.x = element\_line(colour = "black"),

axis.ticks.y = element\_line(colour = "black"),

strip.text.x = element\_text(size = 10)) +

ggtitle("Total Number of Retweets Over Time")

[](https://mathewanalytics.files.wordpress.com/2018/12/content_1.png)

[](https://mathewanalytics.files.wordpress.com/2018/12/content_2.png)

***D2. Frequency of Favourited Tweets***

Crossfitmayhem was the only box where a majority of their tweets were actually favorited at least once. Furthermore, they had a moderately large number of those tweets that were getting favorited three or more times.

## amount of favorited tweets

final\_dat[, .N, by=.(crossfit\_name,favorite\_count)][, percen := N/sum(N), by=.(crossfit\_name)][favorite\_count==0,] %>%

ggplot(., aes(x=reorder(crossfit\_name,percen), y=percen)) +

geom\_bar(stat="identity", fill="steelblue") + coord\_flip() +

labs(x="", y="", colour="") + ylim(0,1) +

theme(plot.title = element\_text(hjust = 0.5),

axis.ticks.x = element\_line(colour = "black"),

axis.ticks.y = element\_line(colour = "black")) +

ggtitle("Percentage of Tweets That Were Not Favorited")

## amount of favorited tweets by count

final\_dat[, .N, by=.(crossfit\_name,favorite\_count)][, percen := N/sum(N), by=.(crossfit\_name)][favorite\_count <= 10,][order(crossfit\_name,favorite\_count)] %>%

ggplot(., aes(x=reorder(favorite\_count, -favorite\_count), y=percen)) +

geom\_bar(stat="identity", fill="steelblue") + coord\_flip() +

labs(x="", y="", colour="") +

facet\_wrap(~crossfit\_name) + ylim(0,1) +

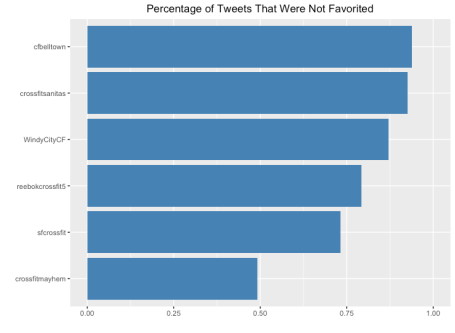
theme(plot.title = element\_text(hjust = 0.5),

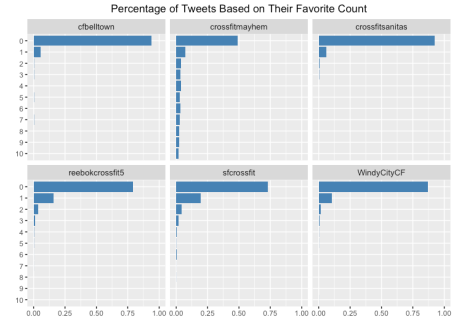
axis.ticks.x = element\_line(colour = "black"),

axis.ticks.y = element\_line(colour = "black"),

strip.text.x = element\_text(size = 10)) +

ggtitle("Percentage of Tweets Based on Their Favorite Count")

[](https://mathewanalytics.files.wordpress.com/2018/12/content_3.png)

[](https://mathewanalytics.files.wordpress.com/2018/12/content_4.png)

***D3. Use of Hashtags***

Reebokcrossfit5 had the largest percentage of tweets that contained hashtags while crossfitsanitas and sfcrossfit rarely used hashtags in their tweets. And when it comes to the hashtags that are being used, it seems that these boxes are using the crossfit hashtag and the name of their own gym.

## contains hashtags

final\_dat[, .N, by=.(crossfit\_name,contains\_hashtags)][, percen := N/sum(N), by=.(crossfit\_name)][contains\_hashtags==1] %>%

ggplot(., aes(x=reorder(crossfit\_name,percen), y=percen)) +

geom\_bar(stat="identity", fill="steelblue") + coord\_flip() +

labs(x="", y="", colour="") + ylim(0,0.8) +

theme(plot.title = element\_text(hjust = 0.5),

axis.ticks.x = element\_line(colour = "black"),

axis.ticks.y = element\_line(colour = "black")) +

ggtitle("Percentage of Tweets That Contain Hashtags")

## frequently used hashtags

final\_dat[hashtags != "", ][,.N, by=.(crossfit\_name, hashtags)][order(crossfit\_name, -N)][,

head(.SD, 3),by=crossfit\_name] %>%

ggplot(., aes(x = reorder(hashtags, -N), y=N)) +

geom\_bar(stat="identity", fill="steelblue") +

coord\_flip() + ylim(0,150) +

facet\_wrap(~crossfit\_name) + labs(x="", y="", colour="") +

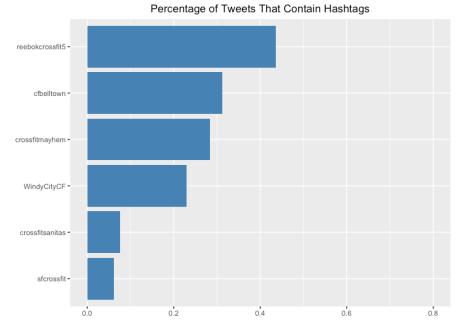
theme(plot.title = element\_text(hjust = 0.5),

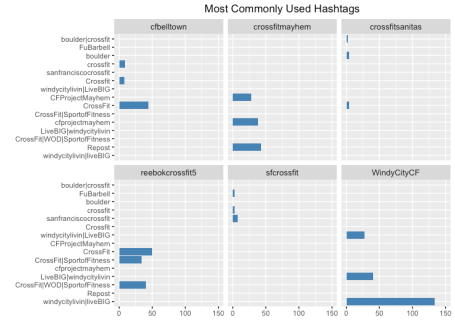
axis.ticks.x = element\_line(colour = "black"),

axis.ticks.y = element\_line(colour = "black"),

strip.text.x = element\_text(size = 10)) +

ggtitle("Most Commonly Used Hashtags")

[](https://mathewanalytics.files.wordpress.com/2018/12/content_5.png)

[](https://mathewanalytics.files.wordpress.com/2018/12/content_7.png)

***D4. Tweet Length***

The following chart shows the summary statistics pertaining to the length of the tweets from each of the crossfit boxes. crossfitsanitas and WindyCityCF seem to have the longest tweets on average, and sfcrossfit has the shortest tweet.

## tweet length

melt(final\_dat[, .(avg\_length = mean(display\_text\_width,na.rm=T),

med\_length = median(display\_text\_width,na.rm=T),

min\_length = min(display\_text\_width,na.rm=T),

max\_length = mean(display\_text\_width,na.rm=T)), by=.(crossfit\_name)], id="crossfit\_name") %>%

ggplot(., aes(x=reorder(crossfit\_name, value), y=value)) +

geom\_bar(stat="identity", fill="steelblue") + coord\_flip() +

labs(x="Date", y="Number of Tweets", colour="Crossfit Box") +

theme(plot.title = element\_text(hjust = 0.5)) +

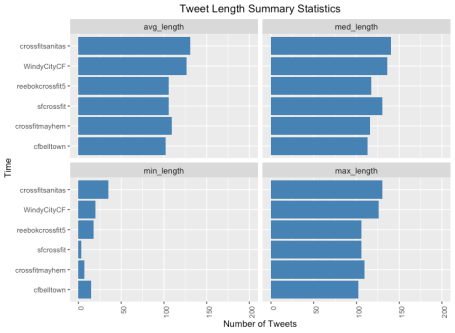
facet\_wrap(~variable) + labs(x="Time", y="Number of Tweets", colour="Crossfit Box") +

theme(axis.text.x=element\_text(angle=90,hjust=1)) +

theme(plot.title = element\_text(hjust = 0.5),

strip.text.x = element\_text(size = 10)) + ylim(0,200) +

ggtitle("Tweet Length Summary Statistics")

[](https://mathewanalytics.files.wordpress.com/2018/12/content_6.png)

**E. Sentiment Analysis**

In order to evaluate the emotion associated with the tweets of each crossfit box, I used the syuzhet package. This package is based on emotion lexicon which maps different words with the various emotions (joy, fear, etc.) and sentiment polarity (positive/negative). We’ll have to calculate the emotion score based on the words present in the tweets and plot the same.

We can see that the majority of tweets for each Crossfit box had a largely positive sentiment, trust and anticipation were other common emotions. For cfbelltown and crossfitsanitas, the third most common emotion was negative.

## sentiment analysis

library(syuzhet)

sentiment\_scores = list()

#each\_gym="sfcrossfit"

for(each\_gym in unique(final\_dat[,crossfit\_name])){

print(each\_gym)

tweet\_text = final\_dat[crossfit\_name==each\_gym, text]

# removing retweets

tweet\_text <- gsub("(RT|via)((?:\\b\\w\*@\\w+)+)","",tweet\_text)

# removing mentions

tweet\_text <- gsub("@\\w+","",tweet\_text)

all\_sentiments <- get\_nrc\_sentiment((tweet\_text))

#head(all\_sentiments)

final\_sentimentscores <- data.table(colSums(all\_sentiments))

#head(final\_sentimentscores)

names(final\_sentimentscores) <- "score"

final\_sentimentscores <- cbind("sentiment"=colnames(all\_sentiments), final\_sentimentscores)

final\_sentimentscores$gym = each\_gym

#final\_sentimentscores

sentiment\_scores[[each\_gym]] <- final\_sentimentscores

}

sentiment\_scores <- rbindlist(sentiment\_scores)

dim(sentiment\_scores)

sentiment\_scores

ggplot(sentiment\_scores, aes(x=sentiment, y=score))+

geom\_bar(aes(fill=sentiment), stat = "identity")+

theme(legend.position="none")+

xlab("Sentiments")+ylab("Scores")+ coord\_flip() +

facet\_wrap(~gym) + labs(x="", y="", colour="") +

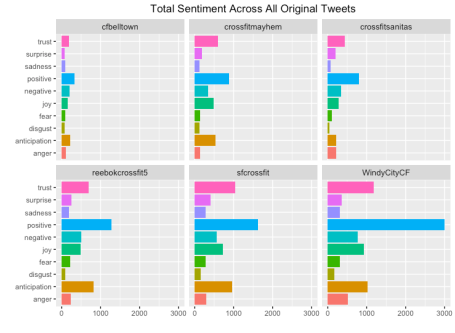
theme(plot.title = element\_text(hjust = 0.5),

axis.ticks.x = element\_line(colour = "black"),

axis.ticks.y = element\_line(colour = "black"),

strip.text.x = element\_text(size = 10)) +

ggtitle("Total Sentiment Across All Original Tweets")

[](https://mathewanalytics.files.wordpress.com/2018/12/sentiment_1.png)

**F. Predictive Model**

Out of curiosity, I wanted to see if a predictive model could be generated to predict whether a tweet would be retweeted. For this task, I looked at just the tweets from sfcrossfit. The target variable was 0 or 1, with zero representing a tweet that was not retweeted and one representing any tweet that was retweeted one or more times. Using the caret package, I trained a random forest on my training data and tried to predict on new data. We can see from the confusion matrix that the model did a very job at predicting whether a tweet would be retweeted or not. Perhaps with some better feature, a better model could have been produced, but I chose not to invest more time on this task.

library(tidyverse)

library(tidytext)

library(stringr)

library(caret)

library(tm)

final\_dat[1:2]

final\_dat[, .N, by=crossfit\_name]

final\_dat\_sub = final\_dat[crossfit\_name=="sfcrossfit",]

final\_dat\_sub[, favorite\_count\_new := ifelse(favorite\_count==0, 0, 1)]

final\_dat\_sub[, retweet\_count\_new := ifelse(retweet\_count==0, 0, 1)]

final\_dat\_sub[1:3]

#final\_dat\_sub$retweet\_count\_new

splt <- createDataPartition(final\_dat\_sub$retweet\_count\_new, p = 0.70, list = F)

train <- final\_dat\_sub[splt, .(created\_at,favorite\_count,is\_retweet, display\_text\_width, contains\_hashtags, retweet\_count\_new)]

dim(train)

test <- final\_dat\_sub[-splt, .(created\_at,favorite\_count,is\_retweet, display\_text\_width, contains\_hashtags, retweet\_count\_new)]

dim(test)

congress\_rf <- train(x = as.matrix(train[,.(year(created\_at), month(created\_at),favorite\_count,is\_retweet, display\_text\_width, contains\_hashtags)]),

y = factor(train$retweet\_count\_new),

method = "rf",

ntree = 200,

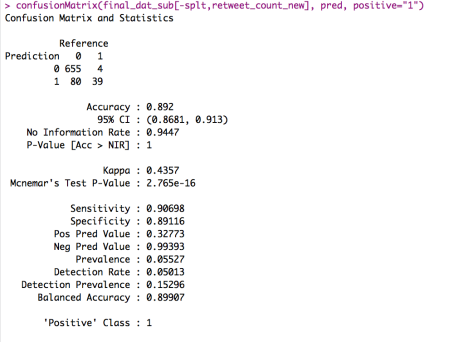
trControl = trainControl(method = "oob"))

congress\_rf$finalModel

varImpPlot(congress\_rf$finalModel)

pred = predict(congress\_rf, as.matrix(final\_dat\_sub[-splt,.(year(created\_at), month(created\_at),favorite\_count,is\_retweet, display\_text\_width, contains\_hashtags)]))

confusionMatrix(final\_dat\_sub[-splt,retweet\_count\_new], pred, positive="1")

[](https://mathewanalytics.files.wordpress.com/2018/12/model_2.png)

**G. Conclusion**

So there you have it. An investigation of how several prominent Crossfit gyms are using their Twitter accounts to engage with consumers. At the end of the day, I would suggest that any business in the health and wellness space should invest more time on Instagram or YouTube than Twitter when it comes to brand marketing or customer acquisition. Twitter is great for news and entertainment, but it isn’t the ideal platform to share fitness content or inspire new members.