Result

Nutthakorn Intharacha

December 18, 2017

Result from analysis

In this reports, I will sum up the result we found from doing analysis for both individual and the team. How are the squads being viwed by the fans and how is the team being recognized by the fans based Twitter. For example if a person says "#Pogba we do not need you, but I love you", this would contribute to #pogba hashtag as 2 points positive ("need", "love"") and 1 point negative ("not).

Individual

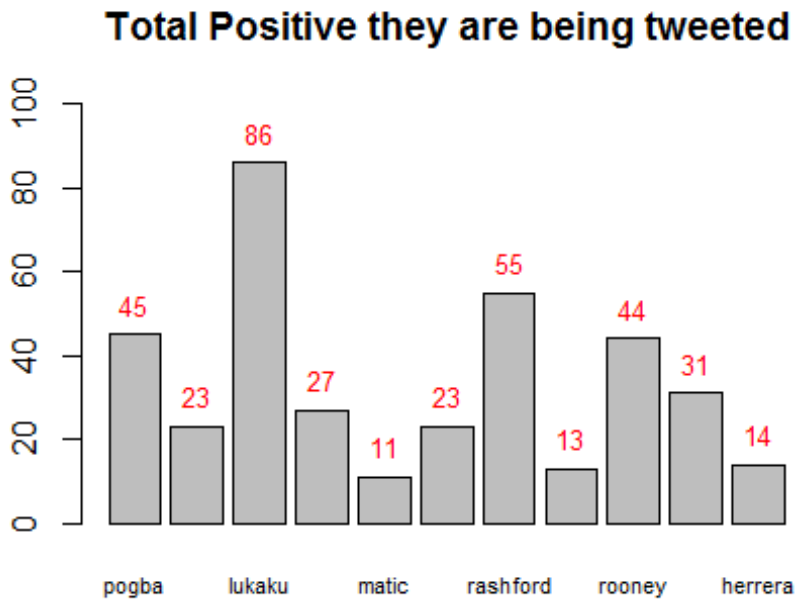
I'd like to start with individual analysis. Here I am going to do Text analysis on the fans' posts on Twitter. Begining ManUtd team's members using Barplot and Piechart for each part-- positive, negative, total, and relative percentage, we will be able to answer these follwing questions 1) Which player is the most being tweeted positively? 2) Which player is the most being tweeted negatively? 3) Which player is being tweeted most frequent? 4) How are their positive images being compared with each other?

At the end of analysis of this part, we will come back to these questions.

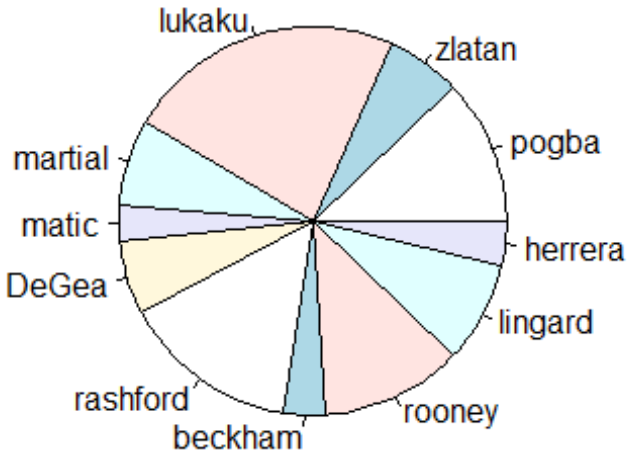
```
load("data_epl_final.RData")
library(ggplot2)
```

Manchester United's player

```
# ManUtd players' barplot
mup1 <- barplot(sscore_manup$positive, names.arg  = manup1, cex.names = 0.7,
main = "Total Positive they are being tweeted", ylim=c(0,100))
text(mup1, y =sscore_manup$positive, label=sscore_manup$positive,
col="red",pos = 3, cex = 0.8)
```

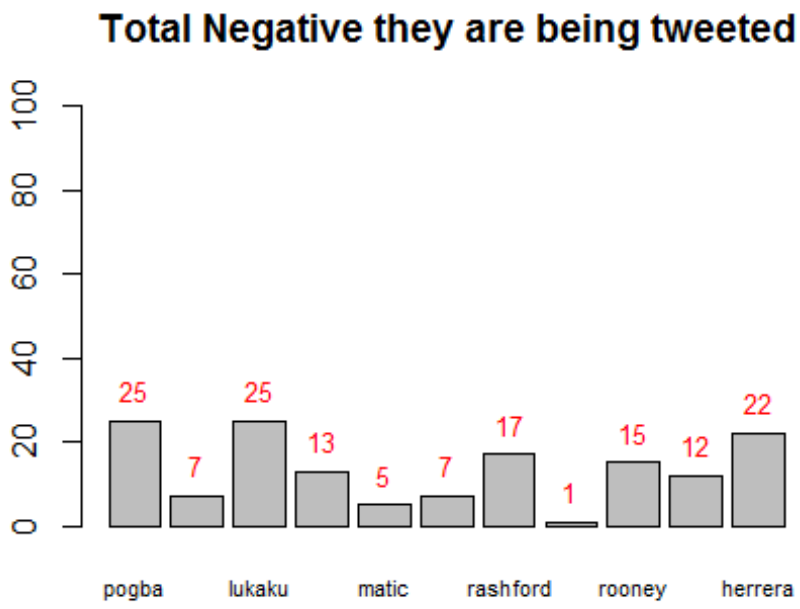



```
pie(sscore_manup$positive,labels = manup1, radius = 1)
```

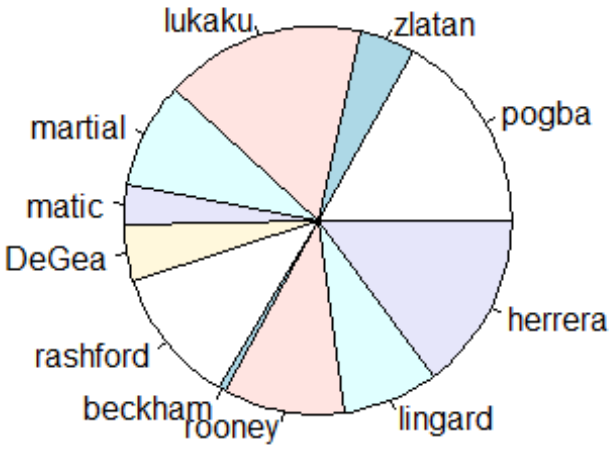


This plot shows the frequency counted for positive words of each person being tweeted. Lukaku, the fresh transfered player from Eventon, unsurprisingly made it to the top one. He scores 14 consecutive games for ManUtd. The second is Rashford, the talented player of the United. And the third is Pogba, the owner of the 80m$ transfer fee. But how are their images, let's look at the nect plot.

```
mup2 <- barplot(sscore_manup$negative, names.arg  = manup1, cex.names =
0.7,main = "Total Negative they are being tweeted", ylim=c(0,100))
text(mup2, y =sscore_manup$negative, label=sscore_manup$negative,
col="red",pos = 3, cex = 0.8)
```



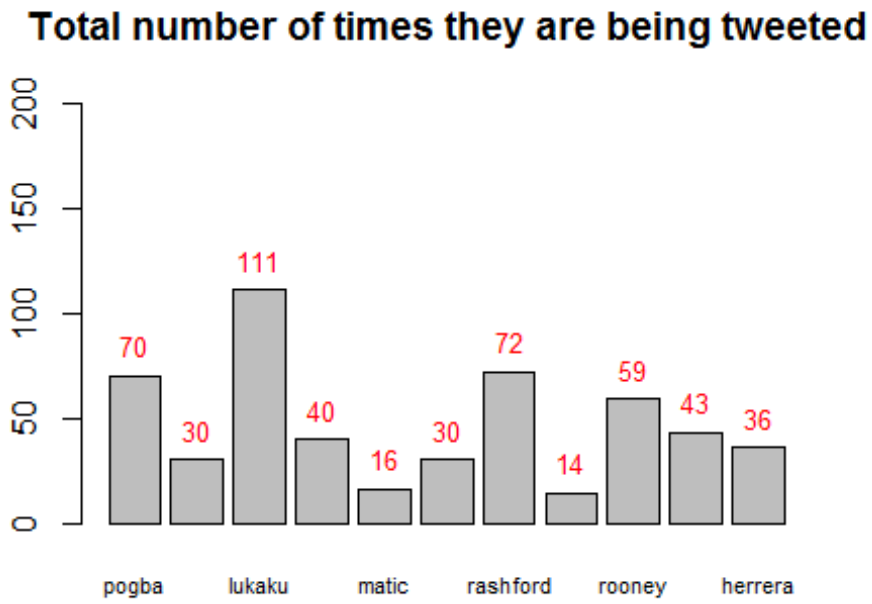

```
pie(sscore_manup$negative,labels = manup1, radius = 1)
```



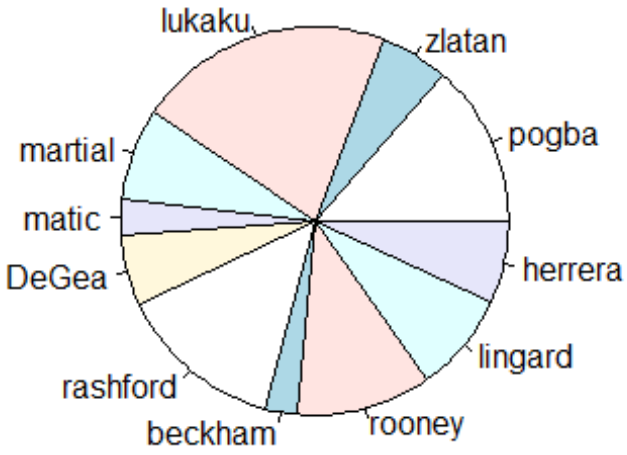
This plot shows how many words they are being tweeted nagtively. YOu see, though Lukaku is number one on the posotive list, but he is also the first one on this list too. Herera suprisingly being negatively tweeted becasue his positive score does not appear much on the first chart. Again Pogba receives many negative comments from the fans. I think it is becasue of his high transfer fee, but his performance does not meet the expectation.

Next, we will look at total number of tweets by the fans, to see which player is being most popular on Twitter regardless of posivitve or negative coments.

```
mup3 <- barplot(sscore_manup$total, names.arg  = manup1, cex.names = 0.7,
main = "Total number of times they are being tweeted", ylim=c(0,200))
text(mup3, y =sscore_manup$total, label=sscore_manup$total, col="red",pos =
3, cex = 0.8)
```

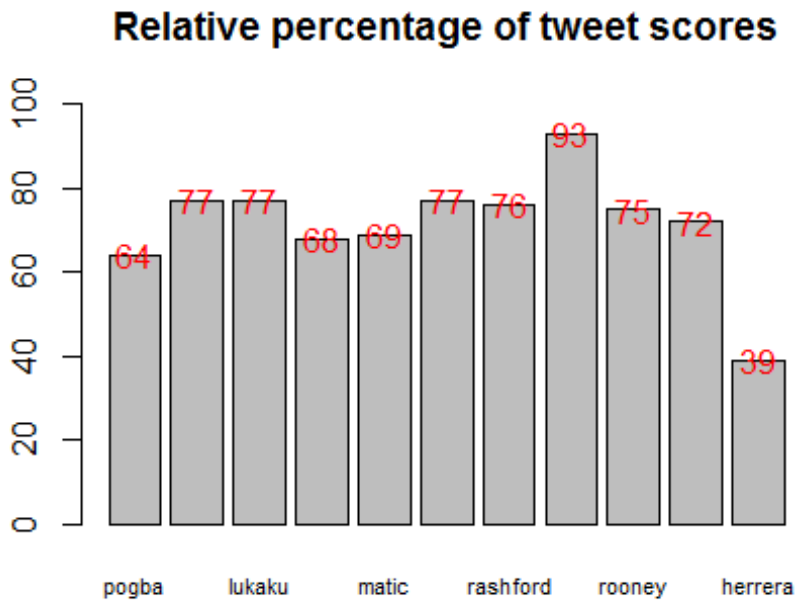



```
pie(sscore_manup$total,labels = manup1, radius = 1)
```

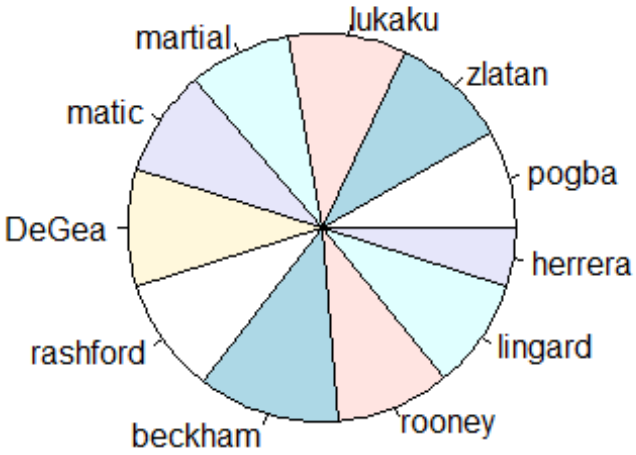


In this plot, Lukeku, without doubt, come with the first place. I assume that there is a big cluster of fans on Twitters who like and dislike him, as well as Pogba. Matic though he is the key man of ManUtd, is not really being tweeted much on Twitter.

```
mup4 <- barplot(sscore_manup$overall, names.arg  = manup1, cex.names = 0.7,
main = "Relative percentage of tweet scores", ylim=c(0,100))
text(mup4, y =sscore_manup$overall, label=sscore_manup$overall, col="red")
```




```
pie(sscore_manup$overall,labels = manup1, radius = 1)
```

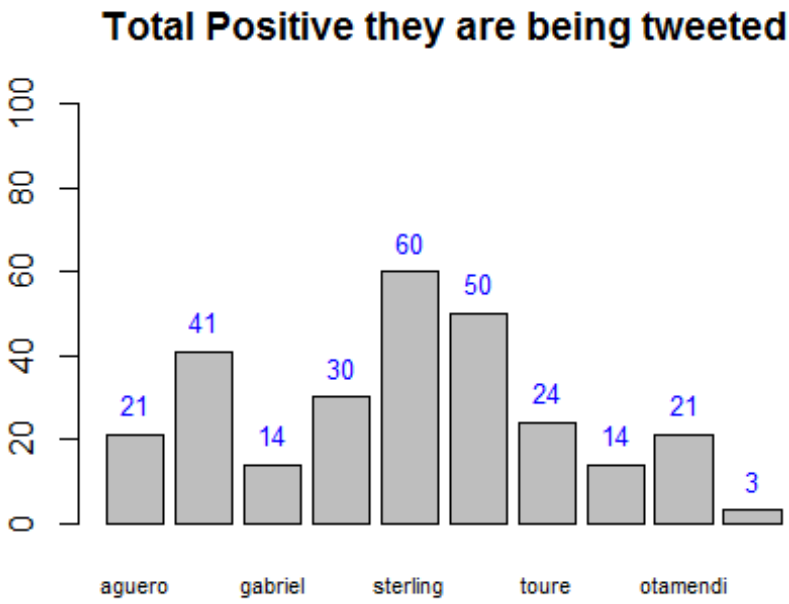


Here are plots for relative percentage of each player. Basically we calculate from Positive scores divided by the total number of words tweeted. It is obvious that most of them are viewed positively by the fans on Twitter as most of them have greater than 50% of relative percentage, except Herrera. The explanation is his performance has been so low recently and there is a news about his transfer to other team.

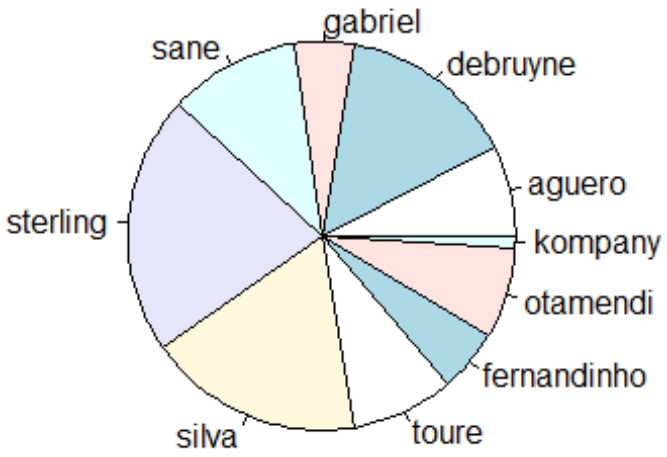
Next we will look at the same analysis for Manchester City.

Manchester City's players

```
mcp1 <- barplot(sscore_mancp$positive, names.arg  = mancp1, cex.names =
0.7,main = "Total Positive they are being tweeted", ylim=c(0,100))
text(mcp1, y =sscore_mancp$positive, label=sscore_mancp$positive,
col="blue",pos = 3, cex = 0.8)
```



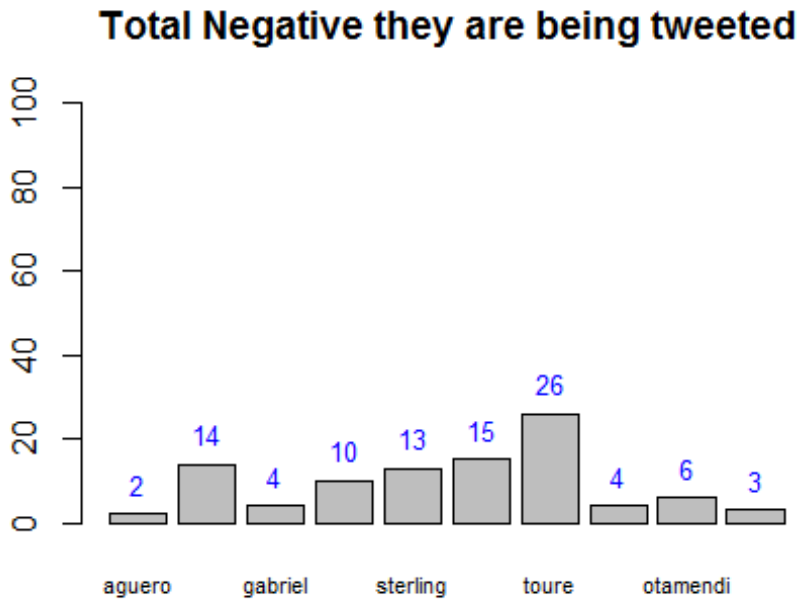

```
pie(sscore_mancp$positive,labels = mancp1, radius = 1)
```



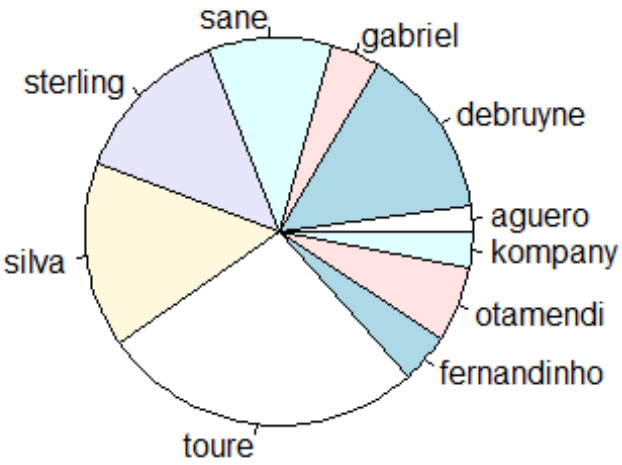
Sterling are the most being positively tweeted...this is a big surprise as he recieve many hates from the fans. However, it is probably becasue ManCity's performance has been outstanding this season with no-losses for home game and all-wins for away games. Silva, the most favorite players voted by the fans in 2016, comes with the second. His form has been getting better and better as well as De Bruyne, the midfielder. Kompany, though is one of the best defenders of English Premier league, come with the last. The possible explanation is that he is getting old.

Just a remark, i think the histrogram looks pretty normal.

```
mcp2 <- barplot(sscore_mancp$negative, names.arg  = mancp1, cex.names =
0.7,main = "Total Negative they are being tweeted", ylim=c(0,100))
text(mcp2, y =sscore_mancp$negative, label=sscore_mancp$negative,
col="blue",pos = 3, cex = 0.8)
```

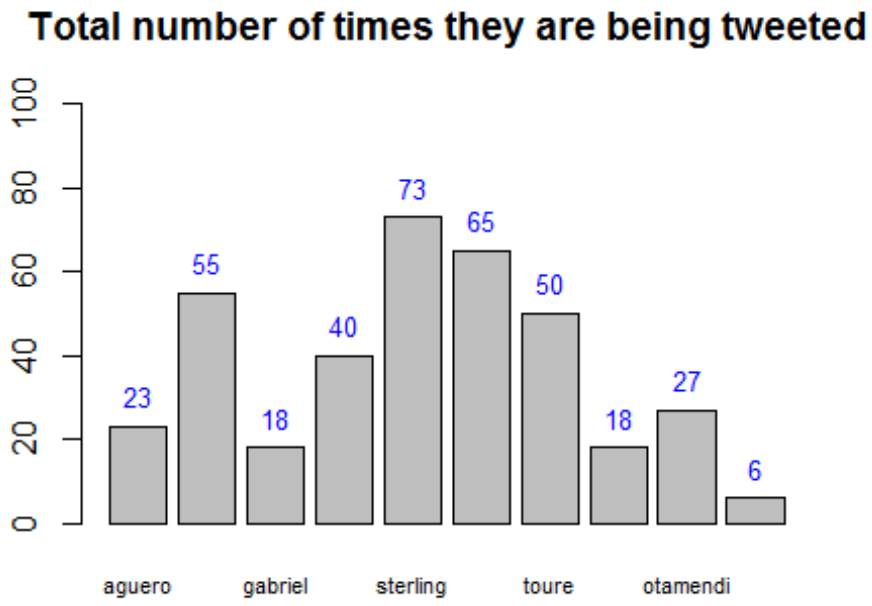



```
pie(sscore_mancp$negative,labels = mancp1, radius = 1)
```

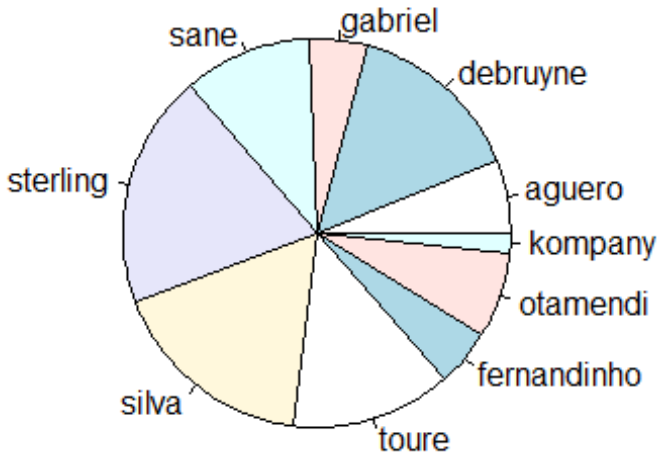


For the negative tweet analysis, Toure is on the top. This make sense becasue he has not played much, but he demands such a high salary, so it is not surprising that the fans might hate him.

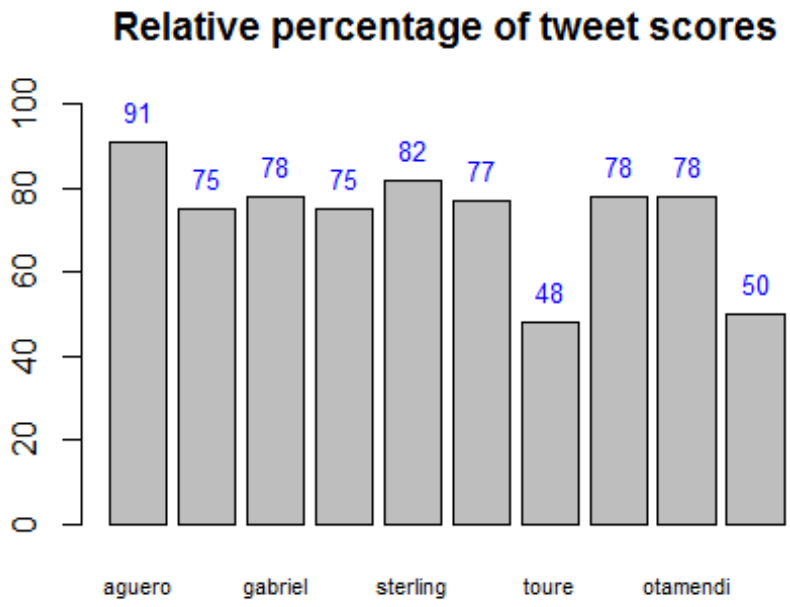
```
mcp3 <- barplot(sscore_mancp$total, names.arg  = mancp1, cex.names = 0.7,
main = "Total number of times they are being tweeted", ylim=c(0,100))
text(mcp3, y =sscore_mancp$total, label=sscore_mancp$total, col="blue",pos =
3, cex = 0.8)
```



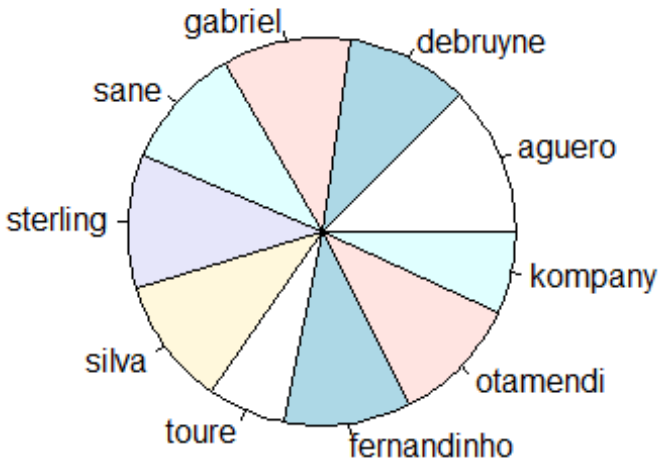

```
pie(sscore_mancp$total,labels = mancp1, radius = 1)
```



```r
mcp4 <- barplot(sscore_mancp$overall, names.arg  = mancp1, cex.names = 0.7,
main = "Relative percentage of tweet scores", ylim=c(0,100))
text(mcp4, y =sscore_mancp$overall, label=sscore_mancp$overall,
col="blue",pos = 3, cex = 0.8)
```




```r
pie(sscore_mancp$overall,labels = mancp1, radius = 1)
```



This part shows that most of the city's players are being loved by the fans. Only Toure, like Herrera from the United, has a relative percentage less than 50%. But for Kompany we cannot really conclude since there are only 6 data points for the analysis.

Now we see that for each team which player is being positively tweeted, negatively tweeted, or the most tweeted on Twitter. We even can roughly tell how are their relative percentage of positive tweets which tells us the public opinions to them in general. Next we want to see the public perception for each team, and if possible can we make a relationship with this individual analysis. i.e is there any relationship such as the higher the score the public have to individuals, the higher the score the public have to the team. We wish to explore this in the next analsyis.

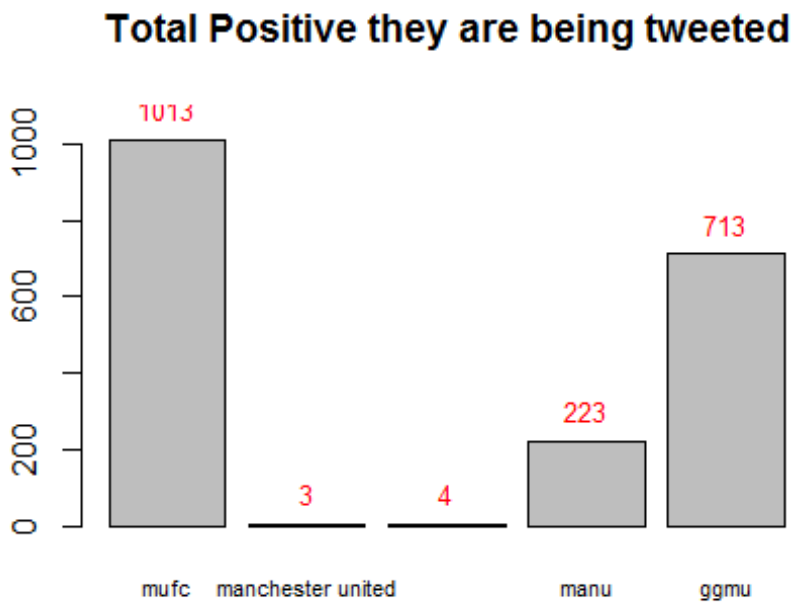
Team

In this section we will explore positive and negative words made by Twitter fans the team. Begining with Manchester United, we use the 5 most hashtaged keyword as you can see in the graph below. We then analyze how many words--both with positive and negative meaning--are made by a person who use these hashtags. For example if A says "#mufc I love you so much" in this case we would cout one as "love" to the hastag "#mufc"

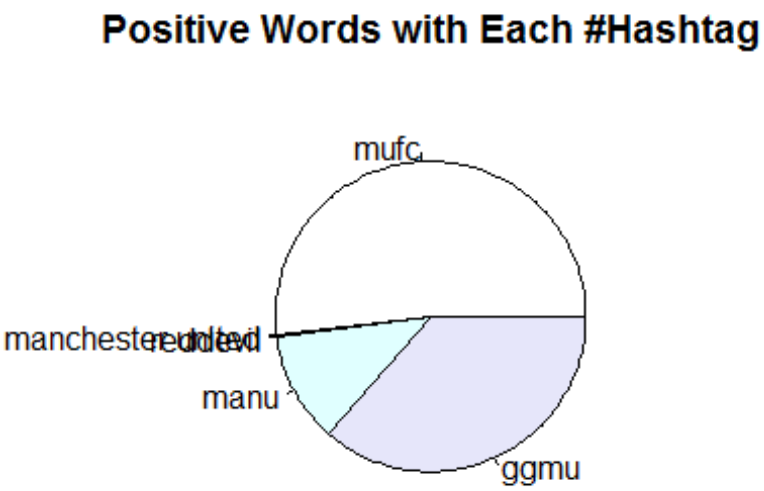
Team: ManUtd

```
# with in a team

mu1 <- barplot(sscore_manu$positive, names.arg  = manu1, cex.names = 0.7,
main = "Total Positive they are being tweeted", ylim=c(0,1100))
```

```
text(mu1, y =sscore_manu$positive, label=sscore_manu$positive, col="red",pos
= 3, cex = 0.8)
```



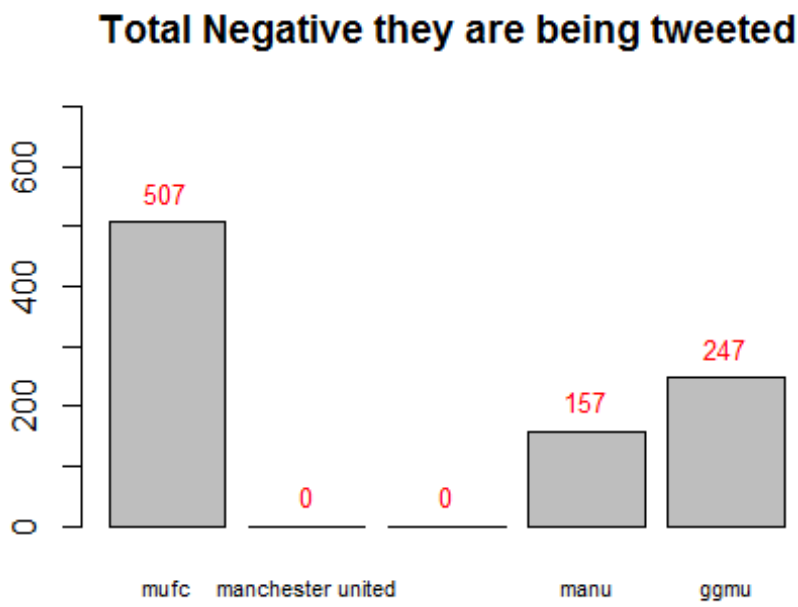

```
# percentage positive
pie(sscore_manu$positive, labels = sscore_manu$team, main="Positive Words
with Each #Hashtag")
```



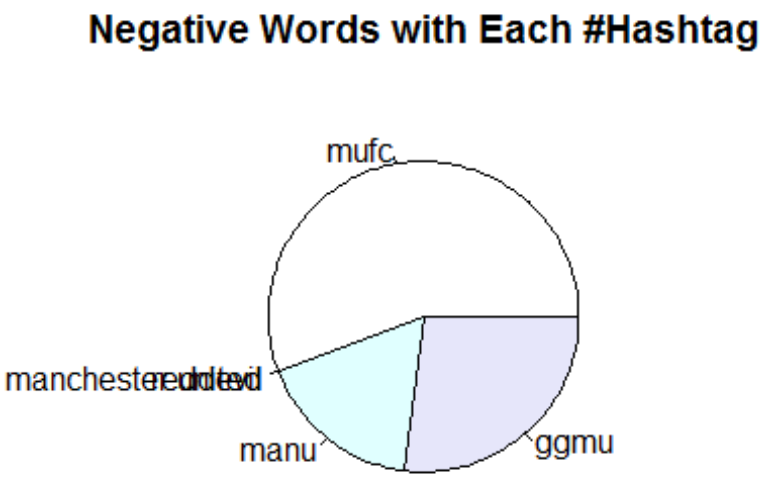
As plot shows, there are 5 hashtags to be analyzed-- #mufc, #manu, #manchsterunited, #ggmu, #redevil.

The hashtag #mufc is apparently dominating the positive comments here. It is understandable becasue people who use #mufc tend to use it with cheerful sentence as well as #ggmu.

```
mu2 <- barplot(sscore_manu$negative, names.arg  = manu1, cex.names = 0.7,main
= "Total Negative they are being tweeted", ylim=c(0,700))
text(mu2, y =sscore_manu$negative, label=sscore_manu$negative, col="red",pos
= 3, cex = 0.8)
```

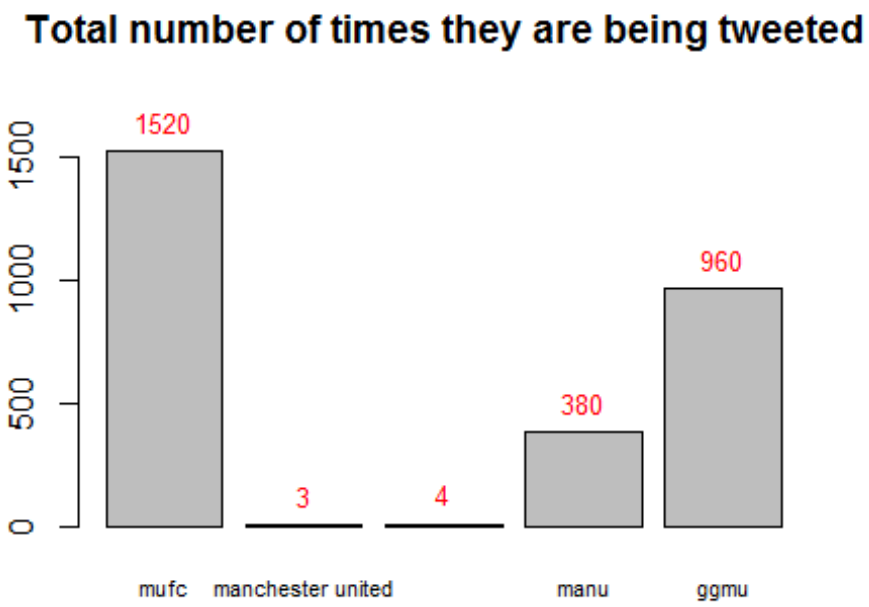



```
# percentage negative
pie(sscore_manu$negative, labels = sscore_manu$team ,main="Negative Words
with Each #Hashtag")
```

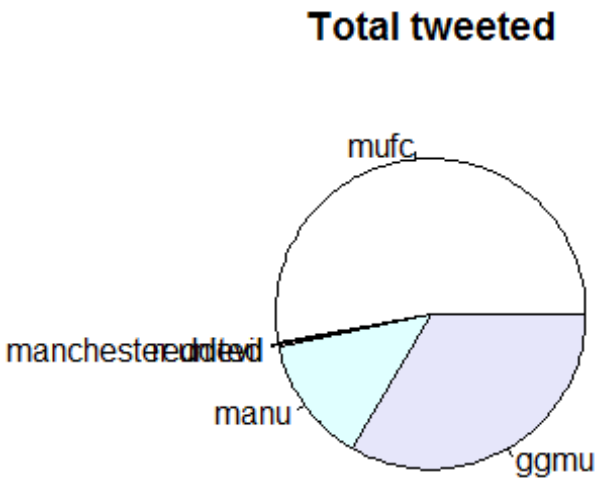


At the same time #mufc is also dominating the negative comments. I think this situaton is when the opponents use #mufc against the United. And the distribution is pretty similar to the positive one. Note that we have few data points for #reddevil and #manchesterunited.

```
mu3 <- barplot(sscore_manu$total, names.arg  = manu1, cex.names = 0.7, main =
"Total number of times they are being tweeted", ylim=c(0,1700))
text(mu3, y =sscore_manu$total, label=sscore_manu$total, col="red",pos = 3,
cex = 0.8)
```

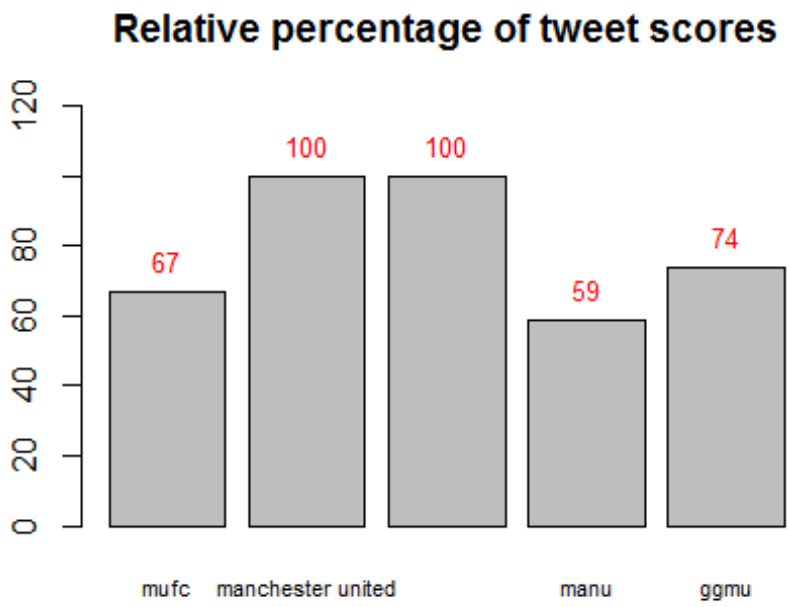



```r
# total
pie(sscore_manu$total, labels = sscore_manu$team, main="Total tweeted")
```

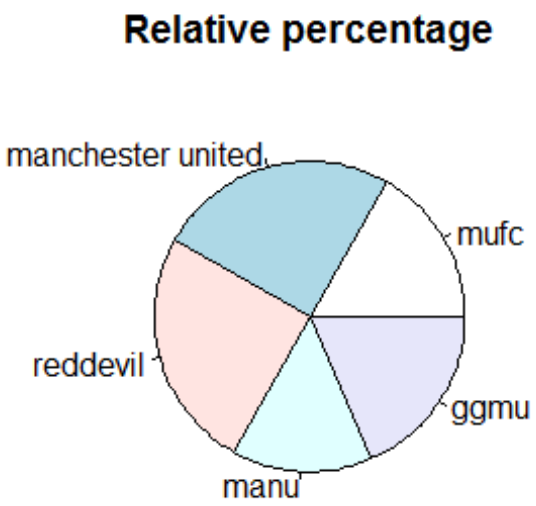


Obviously total number of comments by hashtags top of the list goes to #mufc, following by #ggmu and #manu. Very few #manchesterunited and #reddevil appear on the Twitter.

```
mu4 <- barplot(sscore_manu$overall, names.arg  = manu1, cex.names = 0.7, main
= "Relative percentage of tweet scores", ylim=c(0,120))
text(mu4, y =sscore_manu$overall, label=sscore_manu$overall, col="red",pos =
3, cex = 0.8)
```

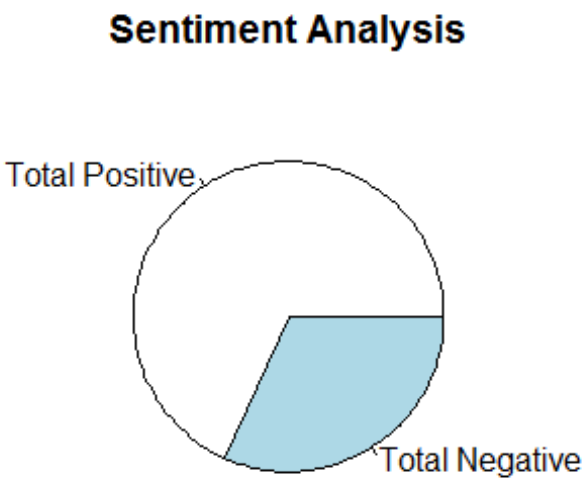



```
pie(sscore_manu$overall, labels = sscore_manu$team, main="Relative
percentage")
```



Here is a relative percentage of positive comment (same method as when we do individual analysis). However, the #manchesterunited and #reddevil are unreliable since we have very few data points.

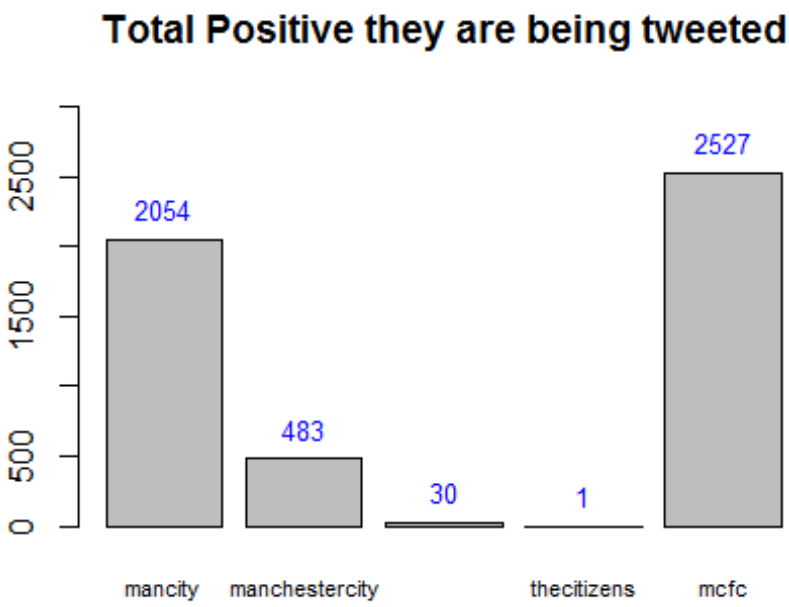
```
# Pos vs neg
manu_sum <- c(manu.ttpos, manu.ttneg)
sent <- c("Total Positive", "Total Negative")
pie(manu_sum, labels = sent, explode=0.0, main="Sentiment Analysis")

## Warning in text.default(1.1 * P$x, 1.1 * P$y, labels[i], xpd = TRUE, adj =
## ifelse(P$x < : "explode" is not a graphical parameter

## Warning in text.default(1.1 * P$x, 1.1 * P$y, labels[i], xpd = TRUE, adj =
## ifelse(P$x < : "explode" is not a graphical parameter

## Warning in title(main = main, ...): "explode" is not a graphical parameter
```



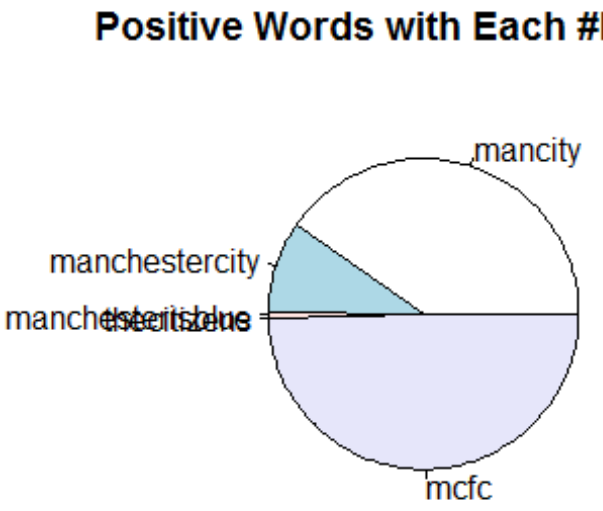
This is total number of positive and negative words regardless the twitter hashtag. You see that the total positive is much greater than total negative. So we can roughly says most fans tweets things positive to the United.

Team ManCity

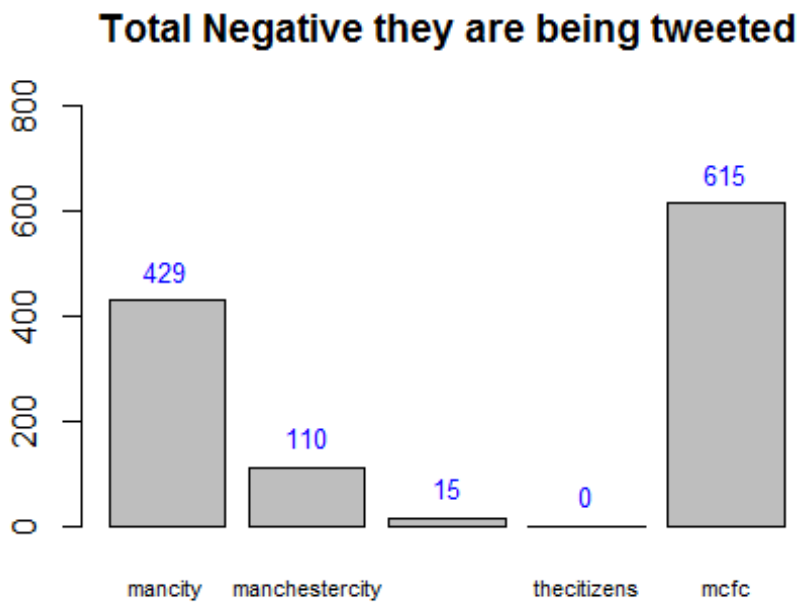
```
mc1 <- barplot(sscore_manc$positive, names.arg  = manc1, cex.names = 0.7,main
= "Total Positive they are being tweeted", ylim=c(0,3000))
text(mc1, y =sscore_manc$positive, label=sscore_manc$positive, col="blue",pos
= 3, cex = 0.8)
```



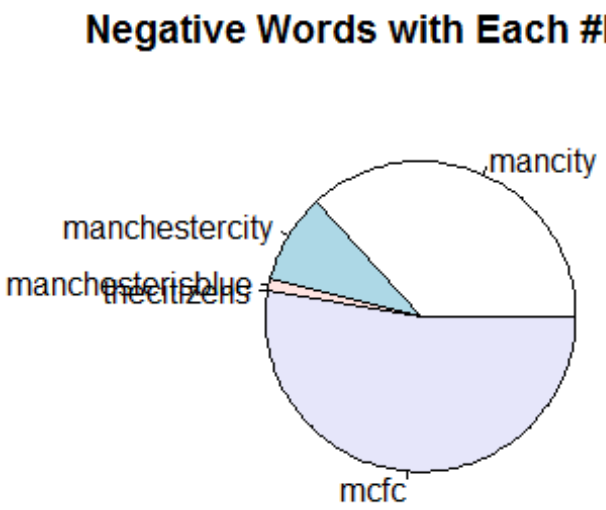

```
# percentage positive
pie(sscore_manc$positive, labels = sscore_manc$team, main="Positive Words
with Each #Hashtag")
```



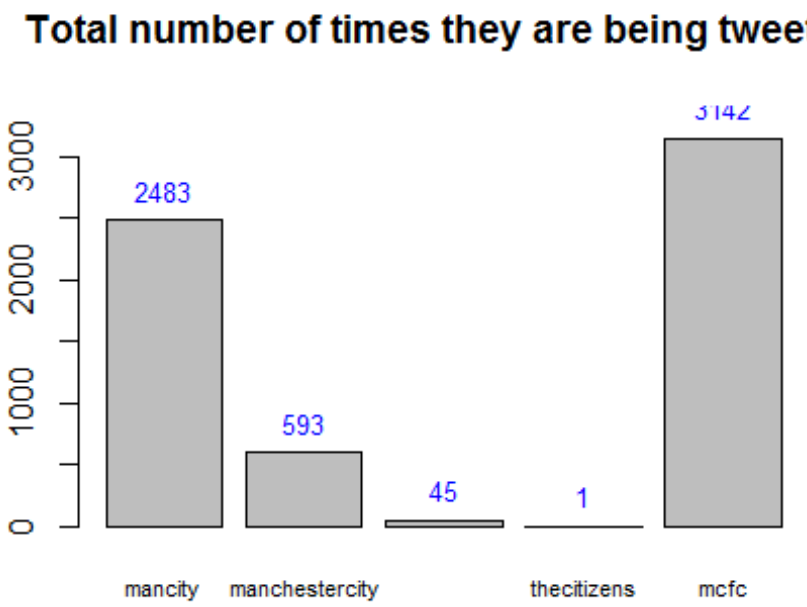
```
mc2 <- barplot(sscore_manc$negative, names.arg  = manc1, cex.names = 0.7,main
= "Total Negative they are being tweeted", ylim=c(0,800))
text(mc2, y =sscore_manc$negative, label=sscore_manc$negative, col="blue",pos
= 3, cex = 0.8)
```



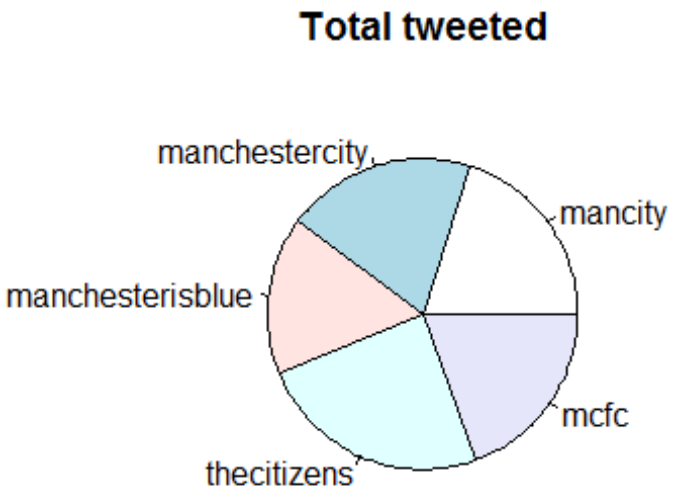

```
pie(sscore_manc$negative, labels = sscore_manc$team ,main="Negative Words
with Each #Hashtag")
```



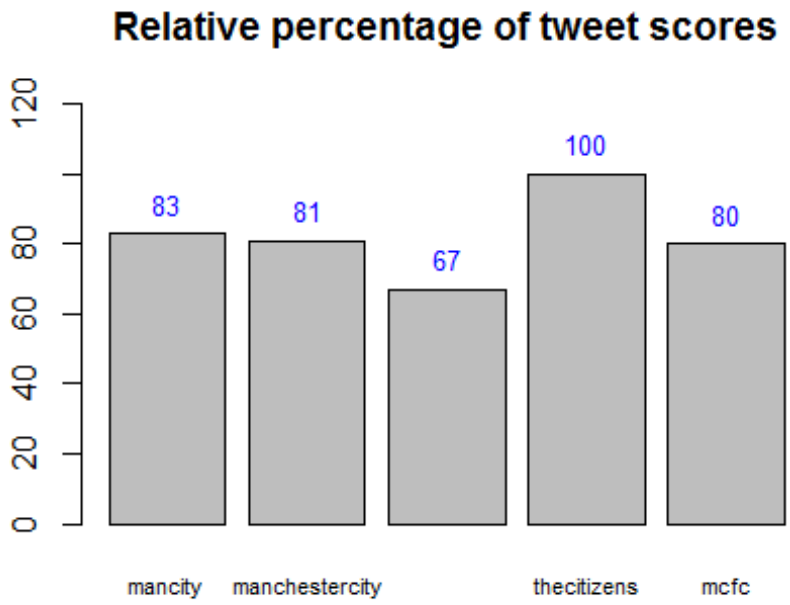

```
mc3 <- barplot(sscore_manc$total, names.arg  = manc1, cex.names = 0.7, main =
"Total number of times they are being tweeted", ylim=c(0,3400))
text(mc3, y =sscore_manc$total, label=sscore_manc$total, col="blue",pos = 3,
cex = 0.8)
```



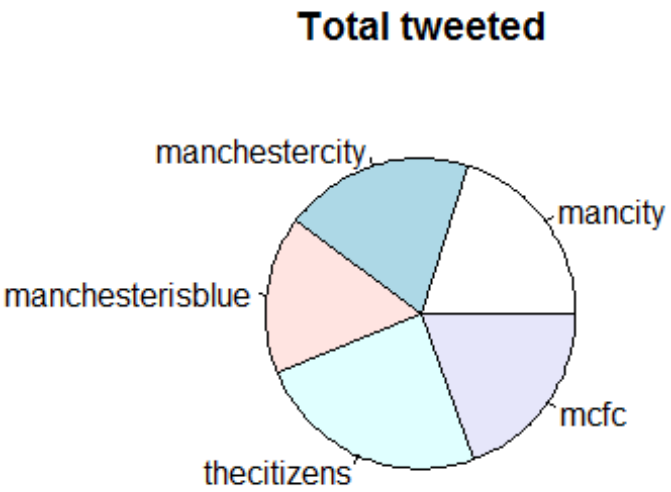

```
pie(sscore_manc$overall, labels = sscore_manc$team, main="Total tweeted")
```



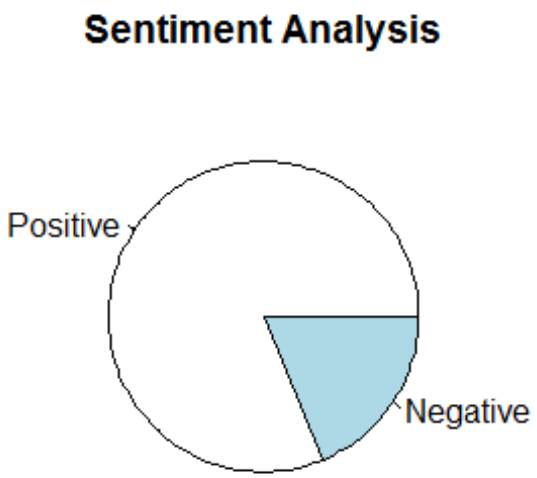
```
mc4 <- barplot(sscore_manc$overall, names.arg  = manc1, cex.names = 0.7, main
= "Relative percentage of tweet scores", ylim=c(0,120))
text(mc4, y =sscore_manc$overall, label=sscore_manc$overall, col="blue",pos =
3, cex = 0.8)
```




```
manc_sum <- c(manc.ttpos, manc.ttneg)
pie(sscore_manc$overall, labels = sscore_manc$team, main="Total tweeted")
```




```
sent <- c("Positive", "Negative")
pie(manc_sum, labels = sent, explode=0.0, main="Sentiment Analysis")

## Warning in text.default(1.1 * P$x, 1.1 * P$y, labels[i], xpd = TRUE, adj =
## ifelse(P$x < : "explode" is not a graphical parameter

## Warning in text.default(1.1 * P$x, 1.1 * P$y, labels[i], xpd = TRUE, adj =
## ifelse(P$x < : "explode" is not a graphical parameter

## Warning in title(main = main, ...): "explode" is not a graphical parameter
```



I want to talk a bit briefly about ManCity team analysis because the logic is the similar. Just want to jump down to the last plot which is a pic chart that shows all positive comments and negative comments regardless the hashtages. It is quite obvious that City fans post more positive words than the United fans. We will look closer in our next analysis, which is Head-to-Head.

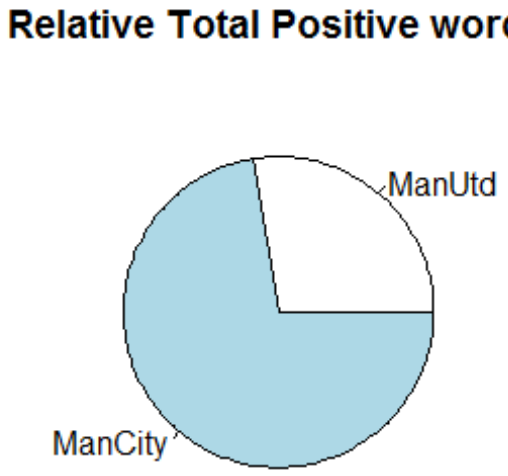
Head to head

This section is where I will implement statistics test compare the positive public opinion of fans to these two team through keywords. We will devided these into three parts 1) Compare positive comments of the fans to these two teams 2) Compare negative comments of the fans to these two teams 3) Compare overall comments of the fans to these two teams (relative percentage) To do so, we will use T-test between two sample means to compare

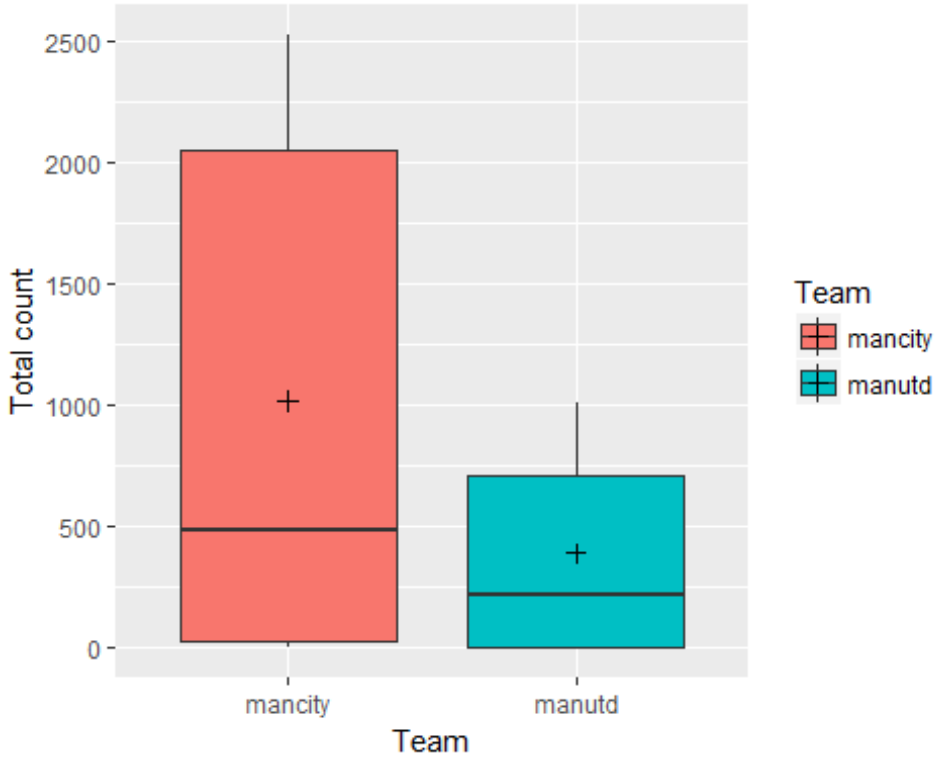
Positive words

```
## pie chart

#mancity vs manu total positive
h2hpos <- c(sumderby[1,2],sumderby[2,2])
sent <- c("ManUtd", "ManCity")
pie(h2hpos, labels = sent, explode=0.0, main="Relative Total Positive words")

## Warning in text.default(1.1 * P$x, 1.1 * P$y, labels[i], xpd = TRUE, adj =
## ifelse(P$x < : "explode" is not a graphical parameter
```

```
## Warning in text.default(1.1 * P$x, 1.1 * P$y, labels[i], xpd = TRUE, adj =
## ifelse(P$x < : "explode" is not a graphical parameter

## Warning in title(main = main, ...): "explode" is not a graphical parameter
```




```
## ggplot
ggplot(derby, aes(x=class, y=positive, fill=class))+
  geom_boxplot()+
  stat_summary(fun.y=mean, geom="point", shape=3, size=2)+
  xlab("Team") + ylab("Total count") +
  guides(fill=guide_legend(title="Team"))
```



This boxplot displays the mean and 95% CI for each team total positive words on the Tweeters. Blue is ManUtd. Red is Mancity. "+" in each box represents mean. We further investigate if their means are differnet. To do so, we construct a t-test

H0: mean "positive" are not different Ha: mean "positive" are different

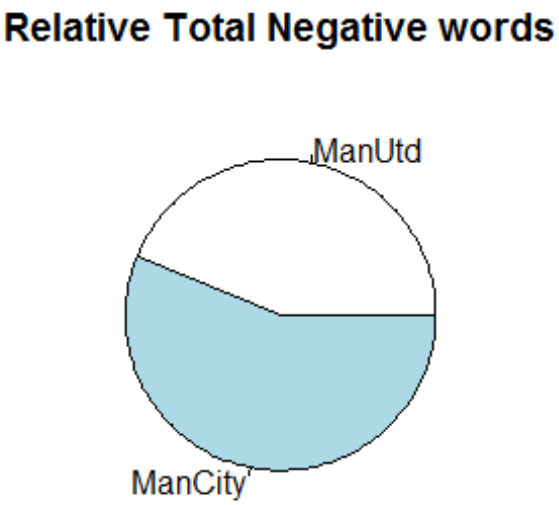
```
testpos <- list()
testpos[[1]] <- t.test(derby$positive~derby$class)
tp <- sapply(testpos,function(x) {
    c(x$estimate[1],
      x$estimate[2],
      ci.lower=x$conf.int[1],
      ci.upper=x$conf.int[2],
      p.value=x$p.value)
})
tp <-  as.data.frame(tp)
colnames(tp) <- c("Stats")

tp
```

```
##                               Stats
## mean in group mancity 1019.0000000
## mean in group manutd   391.2000000
## ci.lower              -822.2292208
## ci.upper              2077.8292208
## p.value                  0.3185625
```

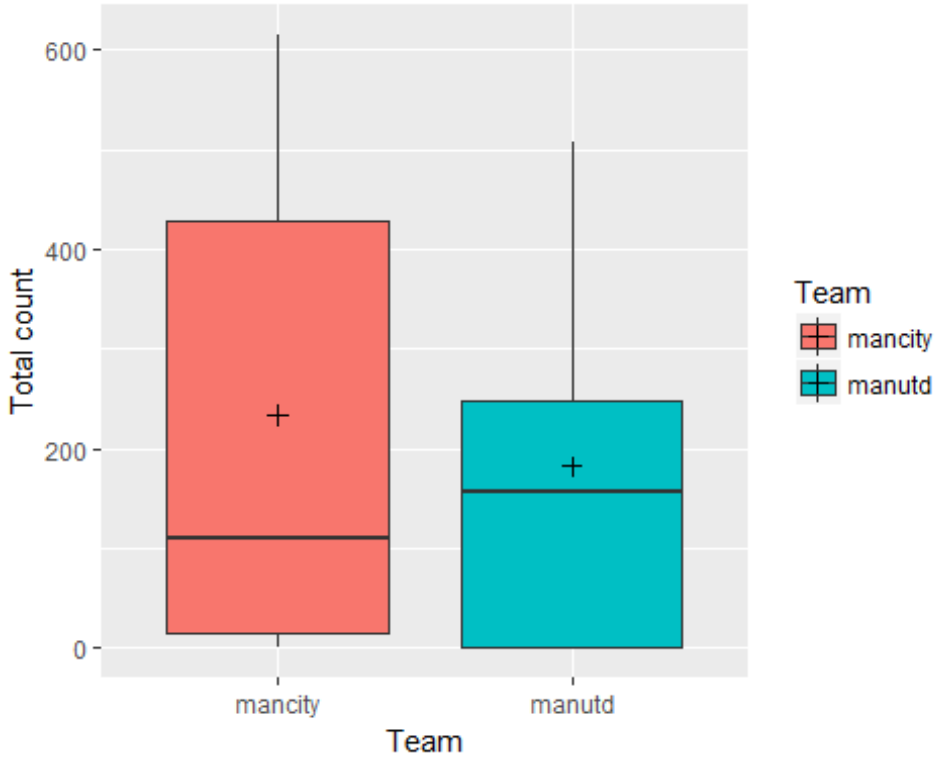
This table show that pvalue = 0.31. So we cannot really say that their positive score means are differnet even though the pie chart and boxplot suggest so.

Negative words

```
h2hneg <- c(sumderby[1,3],sumderby[2,3])
sent <- c("ManUtd", "ManCity")
pie(h2hneg, labels = sent, explode=0.0, main="Relative Total Negative words")

## Warning in text.default(1.1 * P$x, 1.1 * P$y, labels[i], xpd = TRUE, adj =
## ifelse(P$x < : "explode" is not a graphical parameter

## Warning in text.default(1.1 * P$x, 1.1 * P$y, labels[i], xpd = TRUE, adj =
## ifelse(P$x < : "explode" is not a graphical parameter

## Warning in title(main = main, ...): "explode" is not a graphical parameter
```




```
ggplot(derby, aes(x=class, y=negative, fill= class))+
  geom_boxplot() +
  stat_summary(fun.y=mean, geom="point", shape=3, size=2)+
  xlab("Team") + ylab("Total count")+
  guides(fill=guide_legend(title="Team"))
```



As well as the above analysis, we now test of the mean negative tweet scores are different among these two teams.

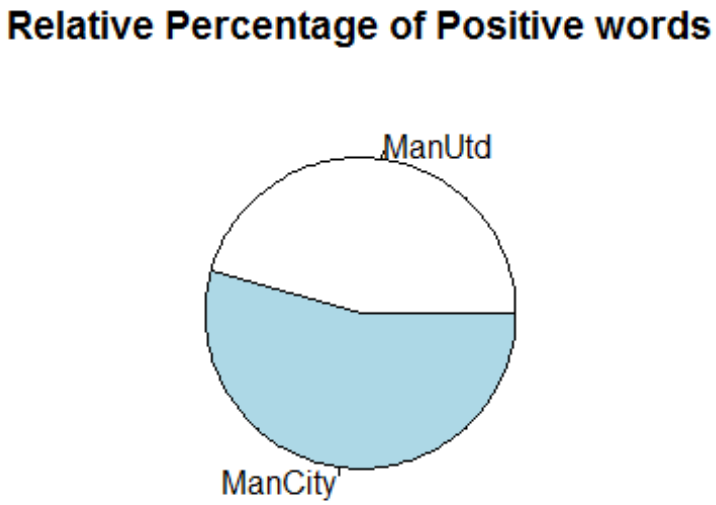
H0: mean "negative" are not different Ha: mean "negative" are different

```
testneg <- list()
testneg[[1]] <- t.test(derby$negative~derby$class)
tn <- sapply(testneg,function(x) {
    c(x$estimate[1],
      x$estimate[2],
      ci.lower=x$conf.int[1],
      ci.upper=x$conf.int[2],
      p.value=x$p.value)
})
tn <-  as.data.frame(tn)
colnames(tn) <- c("Stats")
tn

##                              Stats
## mean in group mancity  233.8000000
## mean in group manutd   182.2000000
## ci.lower              -309.1579182
## ci.upper               412.3579182
## p.value                  0.7477012
```

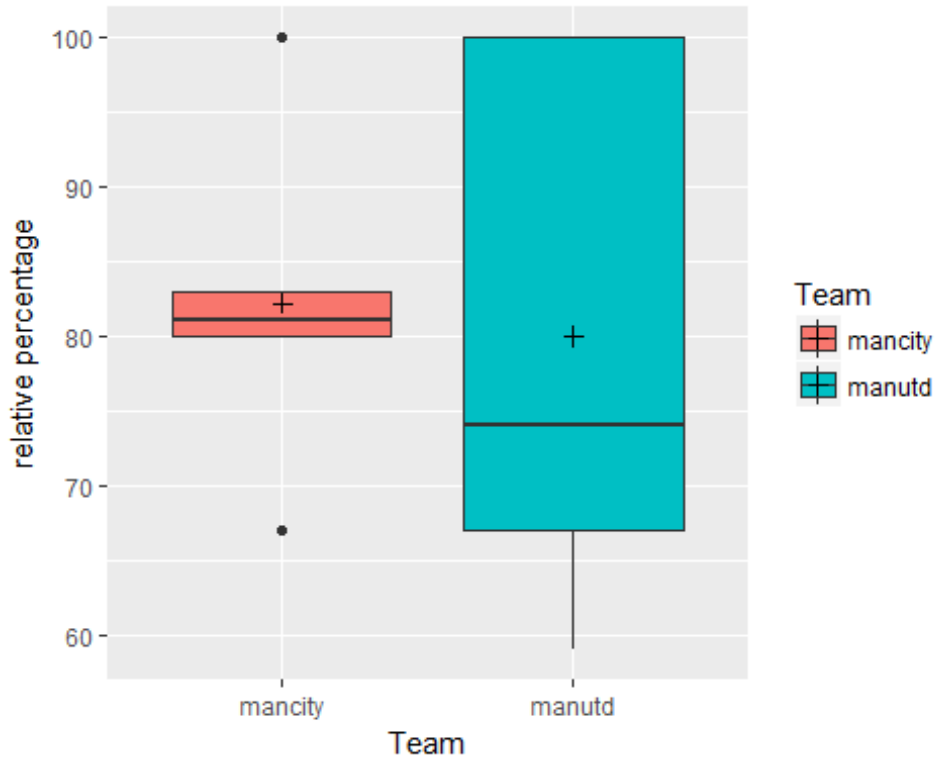
Looking at the table, P-value is 0.7477012 which fail our Ha. Again, no evidence that means are different

Overall relative percentage

```
h2hpct <- c(sumderby[1,4],sumderby[2,4])
sent <- c("ManUtd", "ManCity")
pie(h2hpct, labels = sent, explode=0.0, main="Relative Percentage of Positive
words")

## Warning in text.default(1.1 * P$x, 1.1 * P$y, labels[i], xpd = TRUE, adj =
## ifelse(P$x < : "explode" is not a graphical parameter

## Warning in text.default(1.1 * P$x, 1.1 * P$y, labels[i], xpd = TRUE, adj =
## ifelse(P$x < : "explode" is not a graphical parameter

## Warning in title(main = main, ...): "explode" is not a graphical parameter
```




```
ggplot(derby, aes(x=class, y=overall, fill =class))+
  geom_boxplot()+
  stat_summary(fun.y=mean, geom="point", shape=3, size=2)+
  xlab("Team") + ylab("relative percentage")+
  guides(fill=guide_legend(title="Team"))
```



Lastly, we want to see if the relative percentage of positive tweet are different across these two groups. Recall the relative percentage of positive tweet is calculated by (positive tweets)/(Positive + Negative tweets). Therefore our hypothesis test would be H0: means relative percentages of positive tweet are not different Ha: means relative percentages of positive tweet are different

```
testov <- list()
testov[[1]] <- t.test(derby$overall~derby$class)
tov <- sapply(testov,function(x) {
    c(x$estimate[1],
      x$estimate[2],
      ci.lower=x$conf.int[1],
      ci.upper=x$conf.int[2],
      p.value=x$p.value)
})
tov <-  as.data.frame(tov)
colnames(tov) <- c("Stats")
tov
```

```
##                             Stats
## mean in group mancity  82.2000000
## mean in group manutd   80.0000000
## ci.lower              -21.6859926
## ci.upper               26.0859926
## p.value                 0.8324795
```

Again, P-value =0.83247. We cannot say that the public opinions favor which team than the other.

Conclusion

We have showns three type of analysis--individual, each team, and head-to-head using Twitter based. It is not really surprisng that big name players of each team tend to be tweeted by the fans most frequently regardless positive or negative tweet. We also found that the more positive comments, the more negative comments on that player and on the team. Moreover, it seems like ManUtd, though being tweeted much more frequent than the City, but the proportion of positive and negative words are corresponding making its relative percentage of positive words not as high as we expected. However for the City case, we see a lot of positive words with few negative words to the team. I think maybe their preformance recently being impressive.

Above all, when we look at Head-to-Head analysis using T-test to compare mean difference for each type of words. We found that all three cases--positive, negative, relative percentage--are not really different from each other. At the end of the day, i think amount of people favor ManUtd is as many as amount of those who favor the City, as well as the the amount of people who have negative views toward these two teams.

```
# make summary table
ttesttable <- cbind(c("Mean in group ManUtd","Mean in group ManCity","95% CI
lower for Mean difference","95% CI Upper for Mean difference","P-value for
Mean difference"),tp,tn,tov)
colnames(ttesttable) <- c("Type","positive","negative","overall")
```

Link to my interactive shiny

https://aomnut.shinyapps.io/R415/