Analysis of Facebook Posts

Ali Jawad Fahs, Ashley Lesdalons, Nayanika Dogra and Onss Bousbih

16th January 2017

Abstract

Introduction

Since the creation of social networks such as Facebook and Twitter, the number of users has increased from 1 billion in 2013 to almost 2 billion today, predicting even a bigger increase in the next years. Considering this rapid development, social media may become the most important media channel for brands to reach their clients in the near future.

Realizing the potential of Internet-based social networks, companies use social networks to leverage their businesses. Several studies have been conducted to predict the relation between the online publications on social networks and the impact of such publications measured by users' interactions. We made a similar study to understand the impact of advertisement on Facebook on the number of views and interactions to find out the benefits of putting advertisements of social networks.

The data set studied in this paper is a representative data set of published posts on the Facebook's page of a renowned cosmetic brand between the 1st of January and the 31st of December in 2014. It was found on the UC Irvine Machine Learning Repository. The data set comprises of the following attributes:

- Page Total Likes: The number of likes a page has on all its posts.
- Type: This is the type of the post i.e photo or link.
- Post Month: This is the month in which the post was published.
- Post Weekday: This is the day of the week on which the post was published.
- Post Hour: This is the hour of the day at which the post was published.
- Paid: This tells if the post is paid or non paid.
- Post Viewers: This is the number of users who viewed the post.
- Comment: Number of comments on the post.
- Like: Number of likes on the post.
- Share: Number of times a post was shared.
- Total Interaction: It gives the summation of the comment, like and share.

Using the above data set we verified the following observations:

- Number of views to be expected in case of paid posts.
- Number of views to be expected in case of unpaid posts. The above two observations helped us in finding out the effectiveness of advertising on Facebook.
- Correlation between number of views on a post and number of interactions with the post. This shows that the number of views are directly related to number of interactions.
- Relation between type of a post and number of views and interactions associated with it.
- Relation between time of the post and number of views on the post.

Material and Methods

- Mean: It is the average and is computed as the sum of all the observed outcomes from the sample divided by the total number of events. It is given as:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x$$

- Variance: It is the expectation of the squared deviation of a random variable from its mean, and it informally measures how far a set of (random) numbers are spread out from their mean.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x-\overline{x})^2$$

- Standard Deviation: It is a measure that is used to quantify the amount of variation or dispersion of a set of data values.

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x-\overline{x})^2}$$

- Confidence Interval: It a range of values so defined that there is a specified probability that the value of a parameter lies within it.

$$\overline{x} \pm t^* \frac{s}{\sqrt{n}}$$

- Covariance: It is a measure of the joint variability of two random variables.

$$cov(X,Y) = E[(X-\mu x)(Y-\mu y)]$$

- Correlation: Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together.

$$\rho x, y = corr(X,Y) = \frac{cov(X,Y)}{\rho_x \rho_y} = \frac{E[(X-\mu x)(Y-\mu y)]}{\rho_x \rho_y}$$

Statistical Analysis of the Data

```
paid.data<-read.csv("obj1.csv",header=TRUE)
paid.views<-c(0)
for(i in 1:180)
  paid.views[i]=0
for(i in 1:139)
  paid.views[floor(paid.data$post.viewers.for.paid[i]/1000)]=1+paid.views[floor(paid.data$post.viewers.

df= as.data.frame(cbind(Overall.Cond= 1:180, paid.views))
df
df.freq= as.vector(rep(df$Overall.Cond, df$paid.views))
hist(df.freq,breaks = 300,plot = TRUE,xlim = c(0,50))
axis(side=2, at=c(0,5,10,15,20,25))
axis(side=1, at=c(0:25)*5)
```

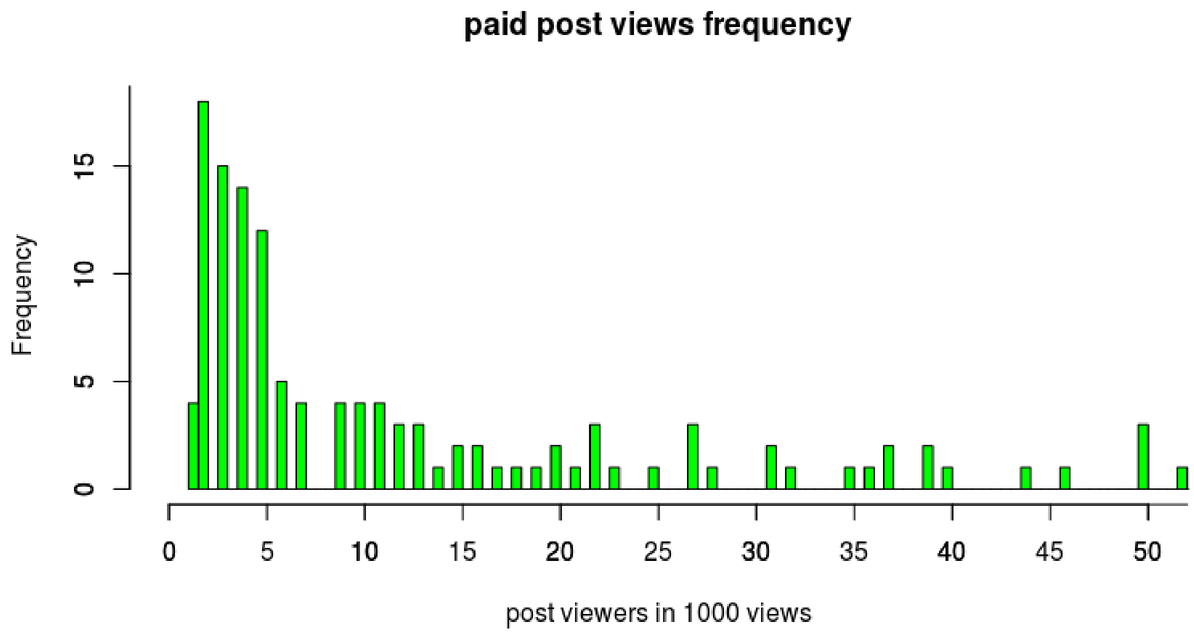


Figure 1:

```
nonpaid.data<-read.csv("obj2.csv",header=TRUE)
nonpaid.views<-c(0)
for(i in 1:180)
  nonpaid.views[i]=0
for(i in 1:361)
  nonpaid.views[floor(nonpaid.data$post.viewers.for.non.paid[i]/1000)]=1+nonpaid.views[floor(nonpaid.da

df= as.data.frame(cbind(Overall.Cond= 1:180, nonpaid.views))
df
df.freq= as.vector(rep(df$Overall.Cond, df$nonpaid.views))
hist(df.freq,breaks = 300,plot = TRUE,col="green",xlim = c(0,50),main="non paid post views frequency",xl
axis(side=2, at=c(0,5,10,15,20,25,30))
axis(side=1, at=c(0:10)*10)
```

```
m1<-mean(paid.data$post.viewers.for.paid)
m1
m2<-mean(nonpaid.data$post.viewers.for.non.paid)
m2

m3<-mean(paid.data$Total.Interactions)
m3
m4<-mean(nonpaid.data$Total.Interactions)
m4

s1<-sd(paid.data$post.viewers.for.paid)
s1
s2<-sd(nonpaid.data$post.viewers.for.non.paid)
```

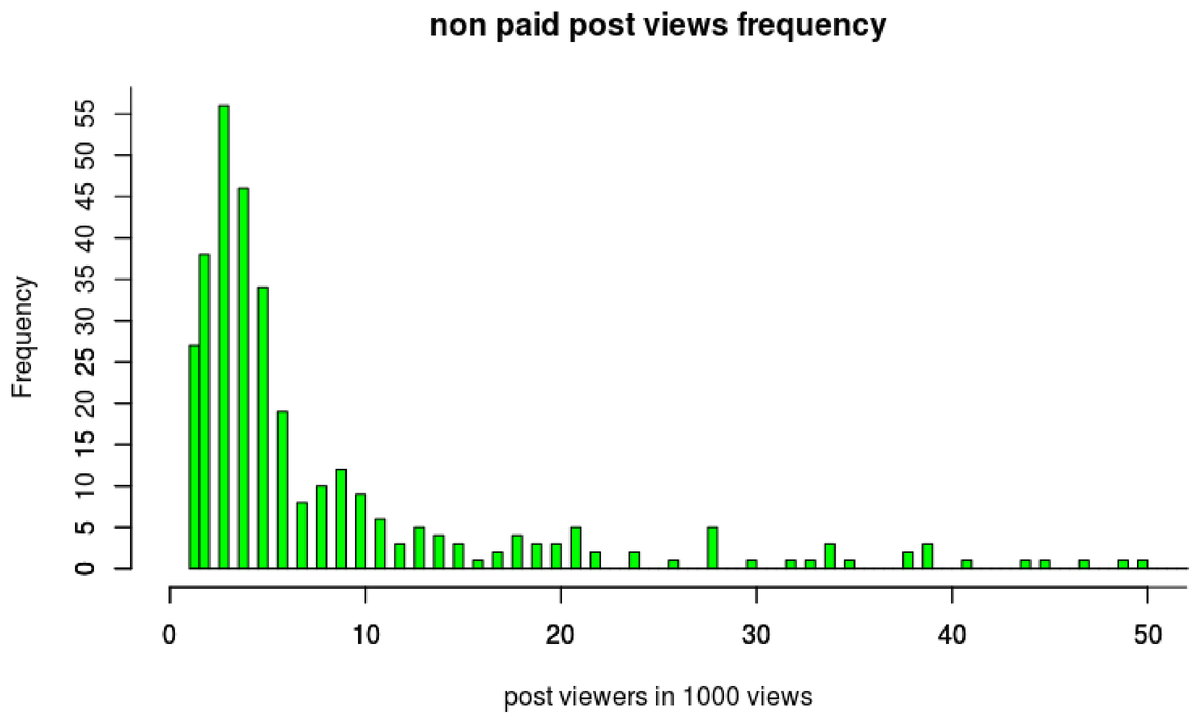
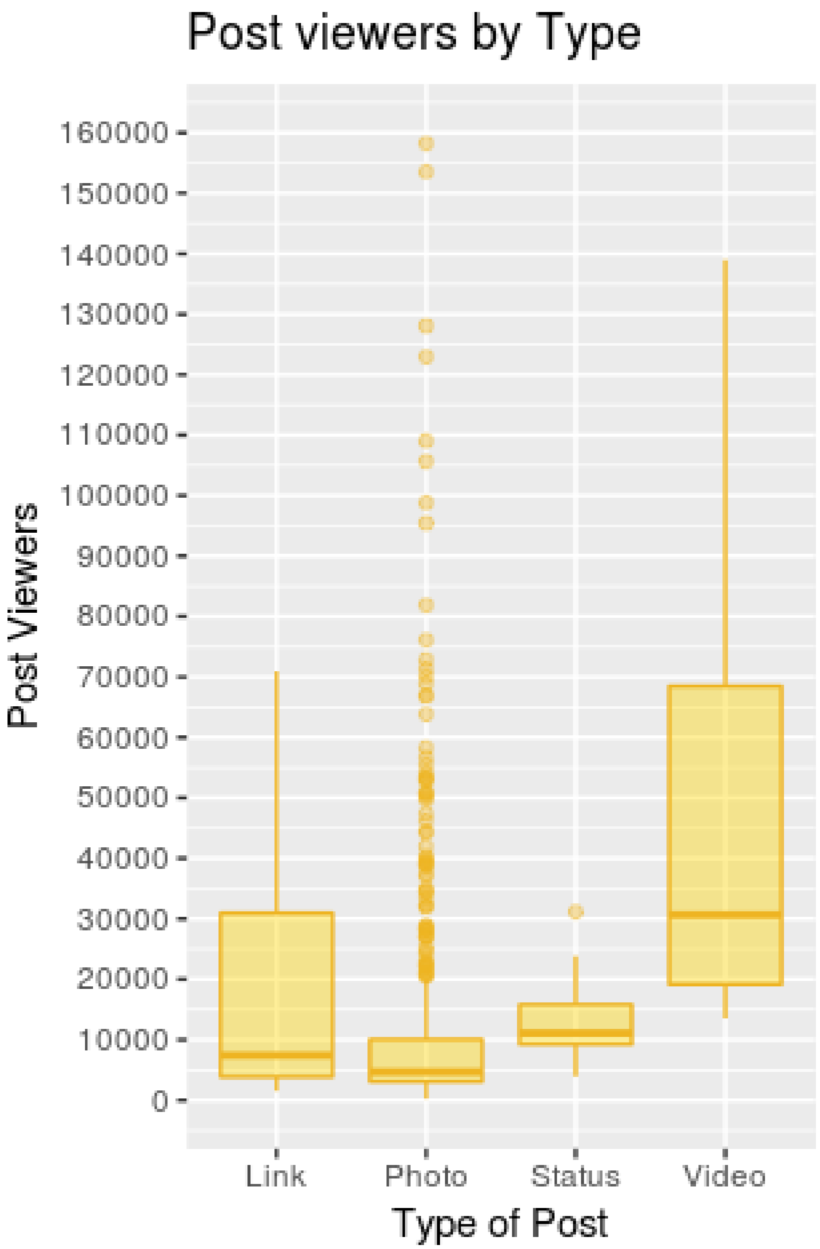
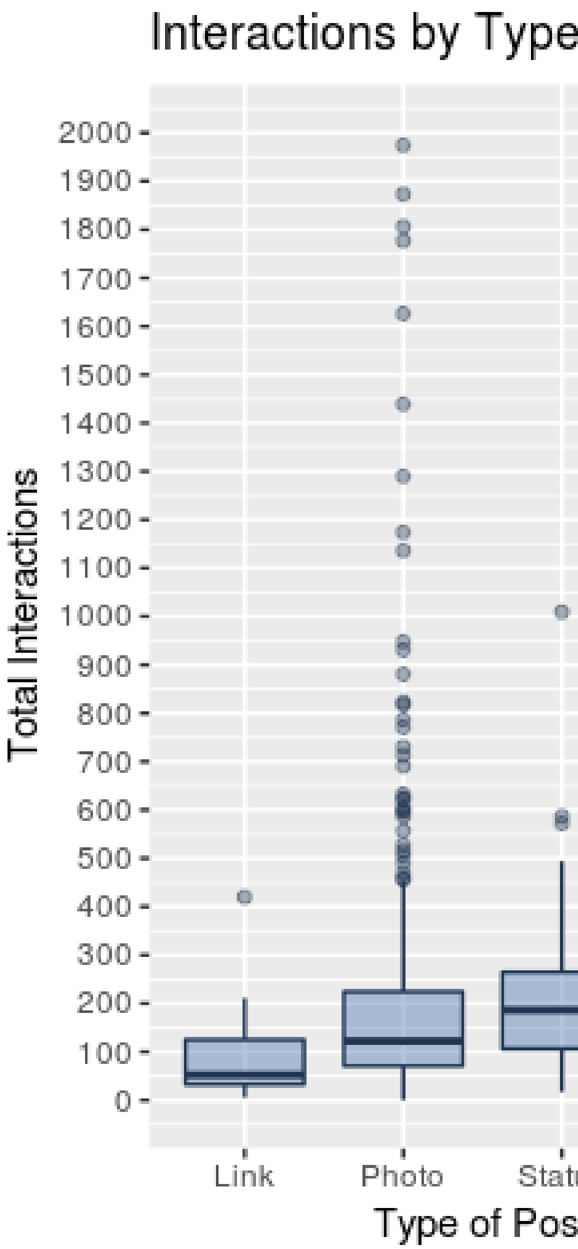


Figure 2:

```
s2

s3<-sd(paid.data$Total.Interactions)
s3
s4<-sd(nonpaid.data$Total.Interactions)
s4

t=1.96
n1=139
n2=361
d1=t*(s1/sqrt(n1))
d2=t*(s2/sqrt(n2))
d1
d2
```

```
MyData3 <- read.csv(file="obj3.csv", header=TRUE)

question3 =ggplot(data = MyData3, aes(MyData3$post.viewers,MyData3$Total.Interactions, color = MyData3$
```





```
MyData4 <- read.csv(file="obj4.csv", header=TRUE)
fill <- "#4271AE"
line <- "#1F3552"

question4 = ggplot(data = MyData4, aes(x = MyData4$Type, y = MyData4$Total.Interactions))+geom_boxplot(

question4

fill <- "gold1"
line <- "goldenrod2"
question4 = ggplot(data = MyData4, aes(x = MyData4$Type, y = MyData4$post.viewers))+geom_boxplot(fill =
```

```
obj5<-read.csv("obj5.csv", header = TRUE)
```

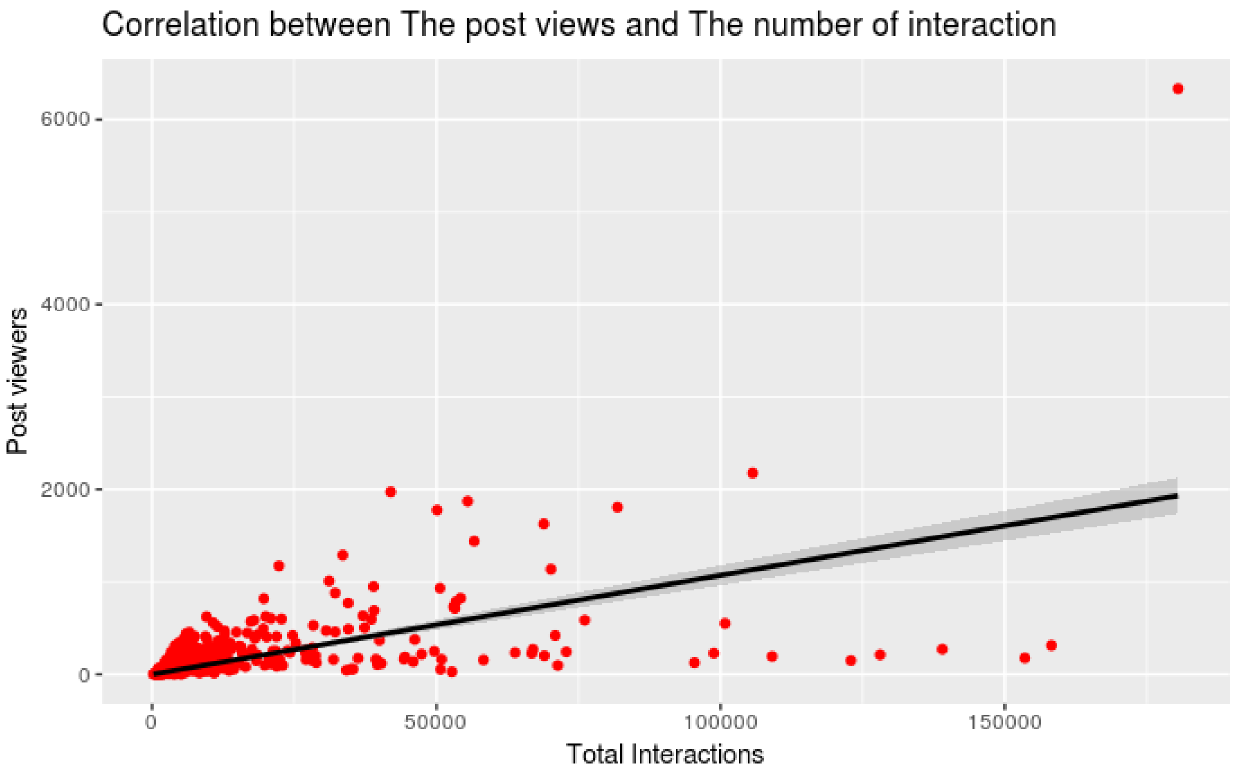


Figure 3:

```
inter=c(0)
views=c(0)
v=c(0)
I=c(0)
for(i in 0:23)
  {v[i]=0
  I[i]=0
  inter[i]=0
  views[i]=0}
for(i in 1:500)
  { v[obj5$Post.Hour[i]]=v[obj5$Post.Hour[i]]+1
I[obj5$Post.Hour[i]]=I[obj5$Post.Hour[i]]+1
 views[obj5$Post.Hour[i]]=views[obj5$Post.Hour[i]]+obj5$post.viewers[i]
inter[obj5$Post.Hour[i]]=inter[obj5$Post.Hour[i]]+obj5$Total.Interactions[i]}

  for(i in 0:23)
  {views[i]=views[i]/v[i]
    inter[i]=inter[i]/I[i]}

df= as.data.frame(cbind(Overall.Cond= 1:23,views))
df
df.freq= as.vector(rep(df$Overall.Cond, df$views))
hist(df.freq,breaks = 24,plot = TRUE,col="green",xlim = c(1,23),main="post views average per one hour i
axis(side=2, at=c(0:8)*5000)
axis(side=1, at=c(0:23))
```

```
df= as.data.frame(cbind(Overall.Cond= 1:23,inter))
df
df.freq= as.vector(rep(df$Overall.Cond, df$inter))
hist(df.freq,breaks = 24,plot = TRUE,col="green",xlim = c(1,23),main="interactions average per one hour
axis(side=2, at=c(0:9)*500)
axis(side=1, at=c(0:23))
```

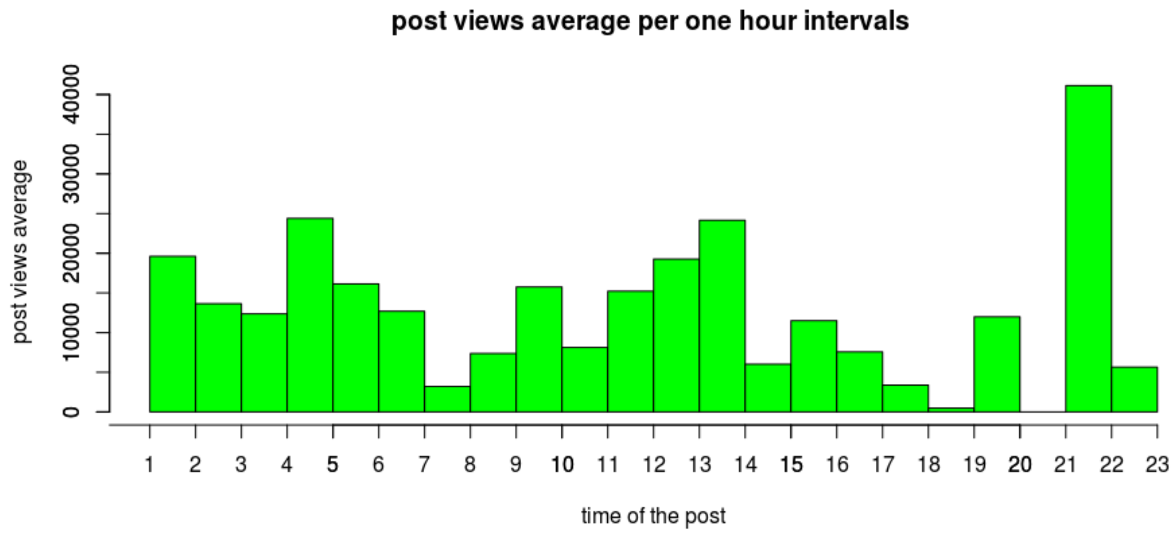
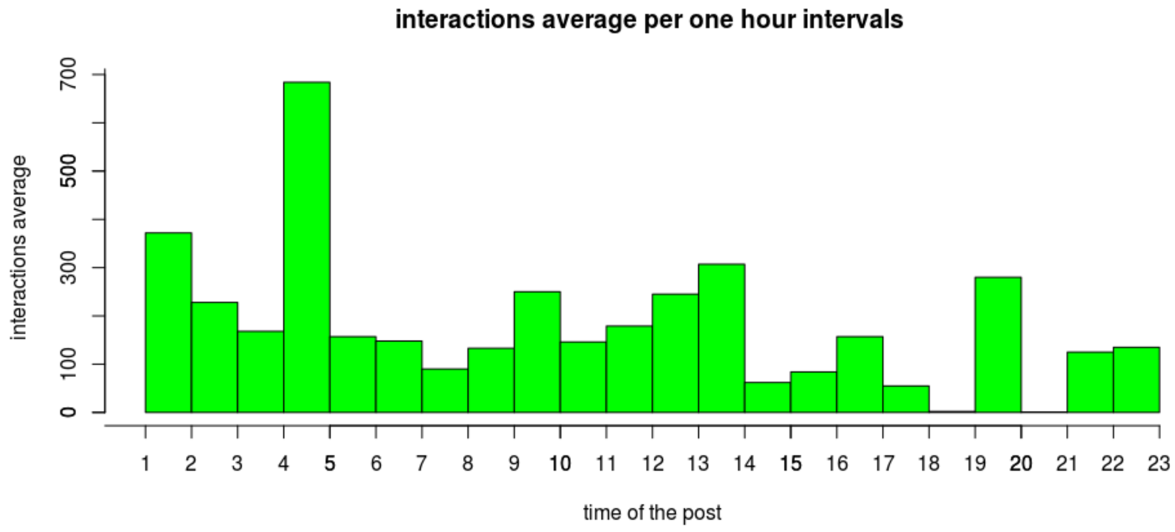


Figure 4:



Conclusion