3 Million Tweet Analysis

January 27, 2019

Integration and cleaning

Integration of Dataset

```
library(data.table)
library(tidyverse)
#install.packages('bit64')

data <-  fread(input =
    "D:/Columbia/Spring 2019/Data Science and Public Policy/Data Assignment 1/IRAhandle_tweets_1.csv")
data$alt_external_id <- as.character(data$alt_external_id)
data$tweet_id <- as.character(data$tweet_id)
for (i in 2:13)
{
  filename <-
  paste("D:/Columbia/Spring 2019/Data Science and Public Policy/Data Assignment 1/IRAhandle_tweets_",
        i,".csv",sep = "")
  data1 <-  fread(input = filename)
  data1$alt_external_id <- as.character(data1$alt_external_id)
  data1$tweet_id <- as.character(data1$tweet_id)
  data <- rbind(data,data1)
}
```

Attributes and size in the dataset

The Attributes in the dataset are

```
names(data)
```

```
##  [1] "external_author_id" "author"             "content"
##  [4] "region"             "language"           "publish_date"
##  [7] "harvested_date"     "following"          "followers"
## [10] "updates"            "post_type"          "account_type"
## [13] "retweet"            "account_category"   "new_june_2018"
## [16] "alt_external_id"    "tweet_id"           "article_url"
## [19] "tco1_step1"         "tco2_step1"         "tco3_step1"
```

The number of tweets in the dataset are

```
nrow(data)
```

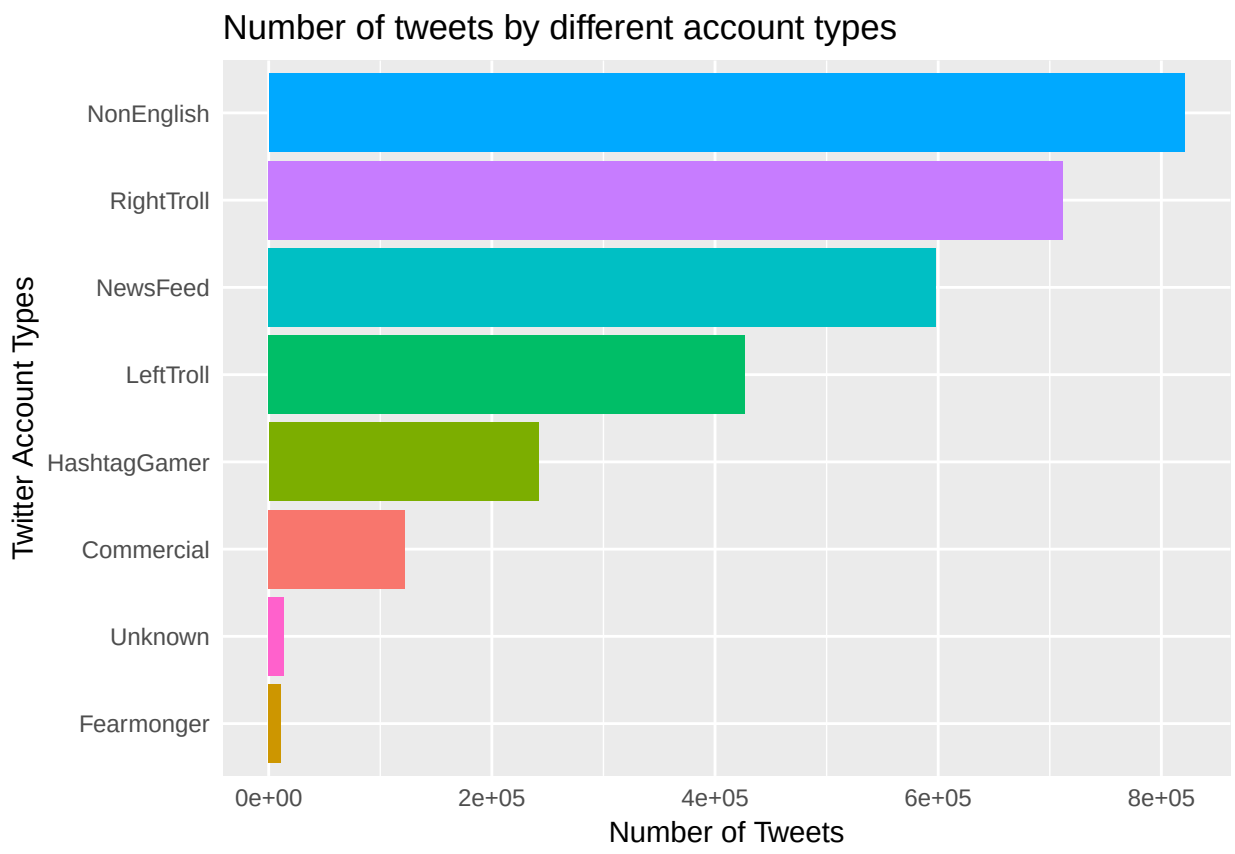
```
## [1] 2946207
```

Preliminary analysis

All Category types

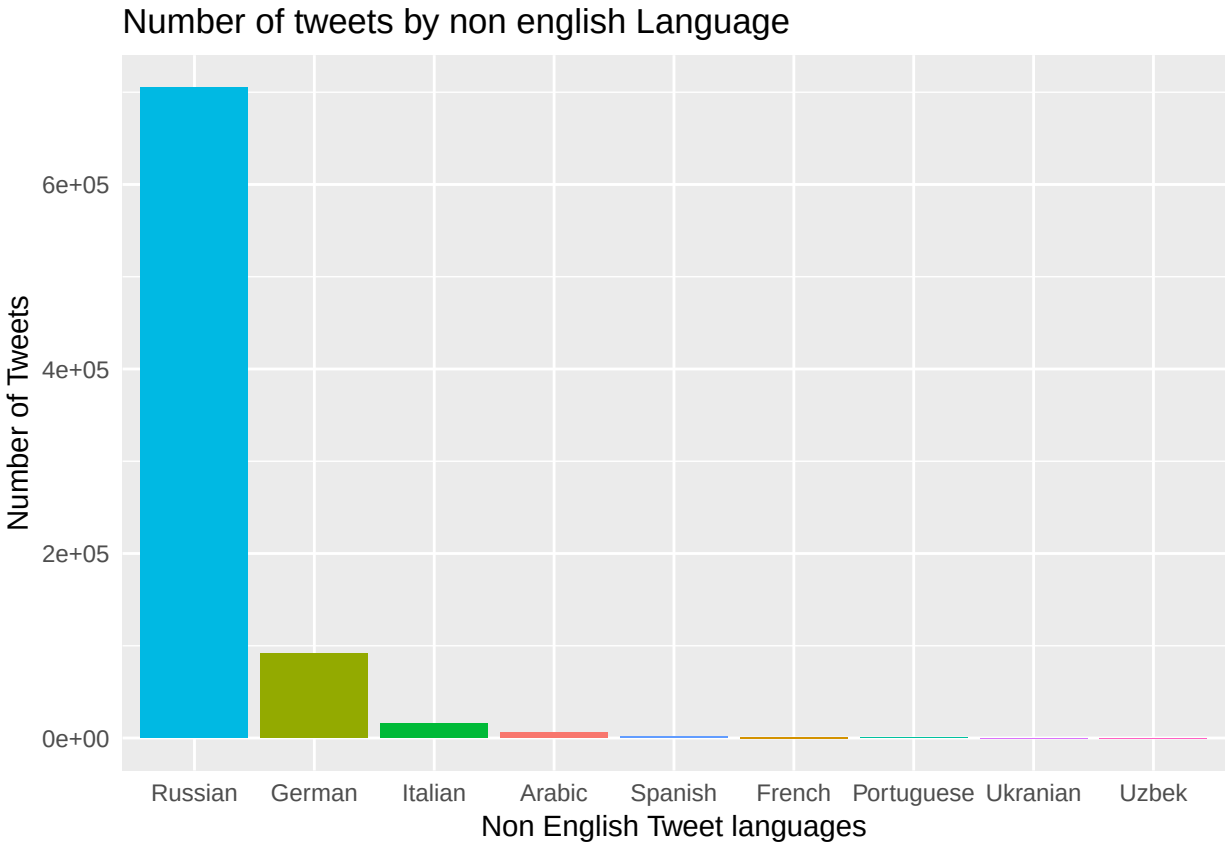
```
# All Category types
acc_cat_data = data[,.(count = .N), by = "account_category"]
```

```
ggplot(acc_cat_data, aes(x = reorder(account_category, count),
                         y = count, fill = account_category)) +
  geom_bar(stat = 'identity') +
theme(axis.ticks = element_blank(),plot.title = element_text(size = rel(1.2)),
      legend.position="none") +
  labs(x = "Twitter Account Types", y = "Number of Tweets",
       title = "Number of tweets by different account types") +
  guides(colour=FALSE) +
  coord_flip()
```



Non English Category

```
# Non English Category
non_eng_category = data[account_category == "NonEnglish",
                        .(count = .N), by = "account_type"]
ggplot(non_eng_category, aes(x = reorder(account_type, desc(count)),
                             y = count, fill = account_type)) +
  geom_bar(stat = 'identity') +
  theme(axis.ticks = element_blank(),plot.title = element_text(size = rel(1.2)),
        legend.position="none") +
  labs(x = "Non English Tweet languages", y = "Number of Tweets",
       title = "Number of tweets by non english Language")+
  guides(colour=FALSE)
```



Retweets Percentages

```
# Retweets Percents
data_counter = data[(account_category == "RightTroll" | account_category == "LeftTroll"),
                    .(count = .N),by = "account_category"]
retweet_counter = data[(account_category == "RightTroll" | account_category == "LeftTroll") &
                         post_type == "RETWEET",.(rtcount = .N),by = "account_category"]

merge.data.frame(data_counter, retweet_counter,account_category=account_category) %>%
  mutate(percent_retweet = rtcount *100 / count)
```

```
##   account_category  count rtcount percent_retweet
## 1        LeftTroll 427141  339475        79.47610
## 2       RightTroll 711668  291609        40.97543
```

Top Retweeters

```
# Top Retweeters
retweeters = head(data[(account_category == "RightTroll" | account_category == "LeftTroll") &
            post_type == "RETWEET",.(rtcount = .N),
            by = c("author", "account_category")][order(rtcount, decreasing = TRUE)],n = 10)
retweeters
```

```
##              author account_category rtcount
##  1:   AMELIEBALDWIN       RightTroll   33678
##  2:  CHESPLAYSCHESS       RightTroll   18597
```

```
##  3: COVFEFENATIONUS       RightTroll   16917
##  4:         HYDDROX       RightTroll   16622
##  5:      ARM_2_ALAN       RightTroll   14551
##  6:      CANNONSHER        LeftTroll   12521
##  7:     ANTONHAYHAY        LeftTroll    9758
##  8:    MRCLYDEPRATT        LeftTroll    9253
##  9: JADONHUTCHINSON        LeftTroll    8439
## 10:    ALECMOOOOODY        LeftTroll    7836
```

Lets analyse AMELIEBALDWIN as the user has most retweets

```
ame_bal_data <- data[author == "AMELIEBALDWIN", .(author, publish_date, content)]
ame_bal_data[,.(count = .N), by = "publish_date"][count>5][order(count, decreasing = TRUE)]
```

```
##         publish_date count
##   1: 9/30/2016 11:30    22
##   2:  10/7/2016 7:37    20
##   3: 9/19/2016 21:08    20
##   4:  10/7/2016 7:49    17
##   5:  10/7/2016 7:50    17
##  ---
## 837: 9/23/2016 14:16     6
## 838: 9/23/2016 15:04     6
## 839: 9/29/2016 16:54     6
## 840: 9/29/2016 19:09     6
## 841: 9/30/2016 11:33     6
```

The number of time AMELIEBALDWIN has tweeted more than 5 tweet a min is 821

Fastest Retweeters

```
# Fastest Retweeters
retweeters = head(data[(account_category == "RightTroll" | account_category == "LeftTroll") &
           post_type == "RETWEET",.(rtcount = .N),
           by = c("author", "account_category")][order(rtcount, decreasing = TRUE)],n = 10)
fastest <- data[(account_category == "RightTroll" | account_category == "LeftTroll") &
                  post_type == "RETWEET", .(author, publish_date, content)]
fastest[,.(count = .N), by = c("publish_date", "author")][order(count, decreasing = TRUE)]
```

```
##            publish_date          author count
##      1:  9/19/2016 7:09   KATERITTERRRR    28
##      2: 9/19/2016 21:07      CASSISHERE    26
##      3:  12/7/2017 9:05 COVFEFENATIONUS    26
##      4: 7/21/2015 20:12  GUILLNAVARRETE    26
##      5: 7/21/2015 22:13      ARM_2_ALAN    25
##     ---
## 243234: 11/11/2015 3:27 _SOLOMONALBERT_     1
## 243235: 12/16/2015 1:17 _SOLOMONALBERT_     1
## 243236:  6/2/2015 20:26 _SOLOMONALBERT_     1
## 243237:  8/9/2015 16:15 _SOLOMONALBERT_     1
## 243238:  9/14/2015 8:07 _SOLOMONALBERT_     1
```

Fastest Orignal Tweeters

```
# Orignal tweeters
f_tweeters = head(data[(account_category == "RightTroll" | account_category == "LeftTroll") &
          post_type == "",.(Retweet_count = .N),
          by = c("author", "account_category")][order(Retweet_count, decreasing = TRUE)],n = 10)
f_tweeters
```

```
##              author account_category Retweet_count
##  1:   WORLDNEWSPOLI       RightTroll         35155
##  2:     JENN_ABRAMS       RightTroll         21169
##  3:       DEBESSTRS       RightTroll         10935
##  4:         TEN_GOP       RightTroll         10388
##  5:     PIGEONTODAY       RightTroll          9026
##  6:       ELIZEESTR       RightTroll          8957
##  7:        CHAASNTR       RightTroll          7781
##  8:  THEFOUNDINGSON       RightTroll          7649
##  9: CRYSTAL1JOHNSON        LeftTroll          7394
## 10:      LAWWAANCTR       RightTroll          6316
```

```
fastest <- data[(account_category == "RightTroll" | account_category == "LeftTroll") &
                  post_type == "", .(author, publish_date, content)]
fast_10 <- fastest[,.(count = .N),
          by = c("publish_date", "author")][count>10][order(count, decreasing = TRUE)]
fast_10
```

```
##          publish_date          author count
##    1: 4/19/2017 10:56 WILLIAMS8KALVIN   116
##    2:  8/4/2017 13:36       ELIZEESTR    26
##    3:  8/16/2017 1:31       DEBESSTRS    25
##    4:  8/13/2017 1:11     MARRISSATRR    24
##    5: 8/12/2017 18:52      ALANISSTRS    23
##   ---
## 3011: 6/12/2017 20:38   WORLDNEWSPOLI    11
## 3012:  7/1/2017 19:50   WORLDNEWSPOLI    11
## 3013:  7/2/2017 14:00   WORLDNEWSPOLI    11
## 3014:  8/1/2017 12:35   WORLDNEWSPOLI    11
## 3015: 8/12/2017 19:34   WORLDNEWSPOLI    11
```

3015 times has more that 10 "Orignal Tweets" been posted per min. WILLIAMS8KALVIN has 116 tweets per min, which makes a case for some of the accounts being bots and not human accounts

```
fast_10[,.(times = .N), by= c("author")][order(times, decreasing = TRUE)]
```

```
##               author times
##   1:       ELIZEESTR   392
##   2:       DEBESSTRS   318
##   3:       ADNNELSTR   186
##   4:      LAWWAANCTR   159
##   5:     MARRISSATRR   119
##  ---
## 118: WILLIAMS8KALVIN     1
## 119:     ANGEELISHET     1
## 120:     BRIISTATRRT     1
## 121:   AMELIEBALDWIN     1
## 122:       SEREESSTT     1
```

There is a possiblility of these 122 accounts being bots

Main Analysis

```
troll_data <- data[(account_category == "RightTroll" | account_category == "LeftTroll"),
                   .(date = as.Date(publish_date, format = "%m/%d/%Y %H:%M"),
                      content, author, account_category)]
```

For the main analysis, we will be considering the Right and left wing trolls and concentrate on the Period during the Wikileaks documnents and The mood of right and left wing trolls. We hypothesize that the Russian trolls were trying to divided the democratic party by cashing in on the wikileaks document leak and trying to display negetive sentiment on Hilary Clinton, while trying to play up Bernie Sanders Supporters by encouraging them to vote against Hillary.

We will be analysing tweets from the period 1st of June 2016 to 31st of August 2016

```
dnc_troll <- troll_data %>%
  subset((date >= as.Date("2016-06-01"))) %>%
  subset((date <= as.Date("2016-08-31")))
```

```
troll_data_counter =
  dnc_troll[(account_category == "RightTroll" | account_category == "LeftTroll"),
                                .(count = .N),by = "account_category"]
troll_data_counter
```

```
##    account_category count
## 1:        LeftTroll 35823
## 2:       RightTroll 16732
```

It is surprising that though the Overall data shows that the Right Troll tweets were about twice of Left Wing Trolls, the left troll accounts were more active in the period of June- August 2016

```
dnc_bernie_left <- dnc_troll[account_category == "LeftTroll" &
                  (grepl("Bernie",content,ignore.case = TRUE) |
                  grepl("Sanders",content,ignore.case = TRUE)),]
dnc_bernie_right <- dnc_troll[account_category == "RightTroll" &
                  (grepl("Bernie",content,ignore.case = TRUE) |
                  grepl("Sanders",content,ignore.case = TRUE)),]
dnc_hillary_left <- dnc_troll[account_category == "LeftTroll" &
                  (grepl("Hillary",content,ignore.case = TRUE) |
                  grepl("Clinton",content,ignore.case = TRUE)),]
dnc_Hillary_right <- dnc_troll[account_category == "RightTroll" &
                  (grepl("Hillary",content,ignore.case = TRUE) |
                  grepl("Clinton",content,ignore.case = TRUE)),]
```

We will be using the R Sentiment analysis package for the analysis. This package provide 3 kinds of sentiment and we will be using SentimentGI, which is based on sentiment on the words of Harvard-IV Dictionary

```
# install.packages("SentimentAnalysis")
library(SentimentAnalysis)
bernie_left_score <- sum(analyzeSentiment(dnc_bernie_left[,content],
                language = "english", aggregate = NULL, removeStopwords = TRUE,
                stemming = TRUE)[2])/nrow(dnc_bernie_left)
bernie_right_score <- sum(analyzeSentiment(dnc_bernie_right[,content],
                    language = "english", aggregate = NULL, removeStopwords = TRUE,
                    stemming = TRUE)[2])/nrow(dnc_bernie_right)
hillary_left_score <- sum(analyzeSentiment(dnc_hillary_left[,content],
                    language = "english", aggregate = NULL, removeStopwords = TRUE,
                    stemming = TRUE)[2])/nrow(dnc_hillary_left)
```

```
hillary_right_score <- sum(analyzeSentiment(dnc_Hillary_right[,content],
                     language = "english", aggregate = NULL, removeStopwords = TRUE,
                     stemming = TRUE)[2])/nrow(dnc_Hillary_right)
cat("Sentiment of Bernie Sanders Among Left Trolls =", bernie_left_score)
```

```
## Sentiment of Bernie Sanders Among Left Trolls = 0.07755223
```

```
cat("Sentiment of Bernie Sanders Among Right Trolls =", bernie_right_score)
```

```
## Sentiment of Bernie Sanders Among Right Trolls = 0.05363493
```

```
cat("Sentiment of Hillary Clinton Among Left Trolls =", hillary_left_score)
```

```
## Sentiment of Hillary Clinton Among Left Trolls = 0.0385431
```

```
cat("Sentiment of Hillary Clinton Among Right Trolls =", hillary_right_score)
```

```
## Sentiment of Hillary Clinton Among Right Trolls = 0.004071132
```

Popular Hashtags in the period

```
right_troll <- dnc_troll[account_category == "RightTroll" & grepl("#", dnc_troll$content)]
left_troll <- dnc_troll[account_category == "LeftTroll" & grepl("#", dnc_troll$content)]

head(setDT(list(unlist(str_extract_all(right_troll$content, "#\\S+"))))[,
                  .(count = .N), by = "V1"][order(count, decreasing = TRUE)], n = 10)
```

```
##                    V1 count
##  1:     #NeverHillary   653
##  2:             #tcot   454
##  3:    #WakeUpAmerica   364
##  4:            #PJNET   352
##  5:   #CrookedHillary   335
##  6:        #Trump2016   305
##  7:               #2A   293
##  8:             #MAGA   206
##  9: #BlackLivesMatter   188
## 10:     #DemsInPhilly   168
```

```
head(setDT(list(unlist(str_extract_all(left_troll$content, "#\\S+"))))[,
                  .(count = .N), by = "V1"][order(count, decreasing = TRUE)], n = 10)
```

```
##                        V1 count
##  1:     #BlackLivesMatter  2160
##  2: #BlackSkinIsNotACrime  1372
##  3:             #StayWoke   406
##  4:        #AltonSterling   401
##  5:     #blacklivesmatter   383
##  6:      #PhilandoCastile   381
##  7:             #staywoke   343
##  8:      #PoliceBrutality   206
##  9:         #MagicButReal   164
## 10:              #Orlando   152
```