Introduction to Tweets Analysis

Analysis of Netflix’s Patriot Act-related Tweets

AbdulMajedRaja RS

20th April 2019

About Me

- Studied at **Government College of Technology, Coimbatore**
- Bengaluru R user group **Organizer**
- R Packages **Developer** (`coinmarketcapr`, `itunesr`)

What's in Twitter for Brands?

When was the last time you

filled a survey

happily

with

full attention&truth?

What's in Twitter for Brands?

- People actually *rant* on Twitter
- **Real** Voice of Customer
- Decent amount of Data

One more **BIG** reason?

One more **BIG** reason?

FREE!!!!

Workflow

Data Collection

- `rtweet`

Data Processing

- `tidyverse`

NLP (Natural Language Processing) & Text Analytics

- `udpipe`
- `tidytext`

Data Visualization

- `ggplot2` (also, part of `tidyverse`)

The Show

rtweet

```r
citation('rtweet')
```

```
##
## To cite rtweet use:
##
##   Kearney, M. W. (2018). rtweet: Collecting Twitter Data. R
##   package version 0.6.7 Retrieved from
##   https://cran.r-project.org/package=rtweet
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{rtweet-package,
##     title = {rtweet: Collecting Twitter Data},
##     author = {Michael W. Kearney},
##     year = {2018},
##     note = {R package version 0.6.7},
##     url = {https://cran.r-project.org/package=rtweet},
##   }
```

Tweet Collection

```r
library(rtweet)

consumer_key ="xxxx"
consumer_secret ="xxxx"
access_token="xxxx"
access_secret="xxxx"


twitter_token = create_token(consumer_key = consumer_key,
                             consumer_secret = consumer_secret,
                             access_token = access_token,
                             access_secret = access_secret)

  keyword1 <- search_tweets('@hasanminhaj india',
                          n = 5000,
                          token = twitter_token,
                          include_rts = FALSE)

write_as_csv(keyword1,
             "~//Documents//R Codes//hasanminhaj_india_noRT.csv")
```

Disclaimer:

- This is a very **naive** Analysis
- **Didn't perform** proper Text Cleaning & Preprocessing, which are very essential
- Objective is to help you get started with **Twitter Analysis**

Loading libraries

```r
library(tidyverse)
library(rtweet)
library(lattice)
library(udpipe)
library(magick)
library(cowplot)
library(ggimage)
library(ggplot2)
library(grid)
library(ggthemes)
```

Data input

```
hasanIN <- read_twitter_csv("hasanminhaj_india_noRT.csv",
                            unflatten = TRUE)
```

```r
# A glimpse of the data
colnames(hasanIN)
```

```
##  [1] "user_id"                 "status_id"
##  [3] "created_at"              "screen_name"
##  [5] "text"                    "source"
##  [7] "display_text_width"      "reply_to_status_id"
##  [9] "reply_to_user_id"        "reply_to_screen_name"
## [11] "is_quote"                "is_retweet"
## [13] "favorite_count"          "retweet_count"
## [15] "hashtags"                "symbols"
## [17] "urls_url"                "urls_t.co"
## [19] "urls_expanded_url"       "media_url"
## [21] "media_t.co"              "media_expanded_url"
## [23] "media_type"              "ext_media_url"
## [25] "ext_media_t.co"          "ext_media_expanded_url"
## [27] "ext_media_type"          "mentions_user_id"
## [29] "mentions_screen_name"    "lang"
## [31] "quoted_status_id"        "quoted_text"
## [33] "quoted_created_at"       "quoted_source"
## [35] "quoted_favorite_count"   "quoted_retweet_count"
## [37] "quoted_user_id"          "quoted_screen_name"
## [39] "quoted_name"             "quoted_followers_count"
## [41] "quoted_friends_count"    "quoted_statuses_count"
## [43] "quoted_location"         "quoted_description"
## [45] "quoted_verified"         "retweet_status_id"
## [47] "retweet_text"            "retweet_created_at"
```

```r
# A glimpse of the data
glimpse(hasanIN)
```

```
## Observations: 1,803
## Variables: 88
## $ user_id                 <chr> "914927933378236416", "99030509067730534…
## $ status_id               <chr> "1108791295131205633", "1108790719131471…
## $ created_at              <chr> "2019-03-21 18:03:39", "2019-03-21 18:01…
## $ screen_name             <chr> "BurntOutCase", "m_complicated_", "aditr…
## $ text                    <chr> "@in_my_sanctuary @PlatinumJab @hasanmin…
## $ source                  <chr> "Twitter for Android", "Twitter for Andr…
## $ display_text_width      <int> 256, 195, 148, 234, 134, 262, 186, 280, …
## $ reply_to_status_id      <chr> "1108789333052608513", "1108789349322248…
## $ reply_to_user_id        <chr> "914927933378236416", "1950140599", NA, …
## $ reply_to_screen_name    <chr> "BurntOutCase", "ItsGazab", NA, "Netflix…
## $ is_quote                <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE…
## $ is_retweet              <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE…
## $ favorite_count          <int> 0, 1, 3, 0, 0, 5, 0, 0, 0, 0, 0, 0, 10, …
## $ retweet_count           <int> 0, 0, 0, 0, 0, 2, 0, 1, 0, 0, 0, 0, 7, 6…
## $ hashtags                <list> [NA, NA, NA, NA, NA, "PatriotAct", NA, …
## $ symbols                 <list> [NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ urls_url                <list> [NA, NA, NA, NA, NA, NA, NA, "twitter.c…
## $ urls_t.co               <list> [NA, NA, NA, NA, NA, NA, NA, "https://t…
## $ urls_expanded_url       <list> [NA, NA, NA, NA, NA, NA, NA, "https://t…
## $ media_url               <list> [NA, NA, "http://pbs.twimg.com/media/D2…
## $ media_t.co              <list> [NA, NA, "https://t.co/Ef7BtDBUWq", NA,…
## $ media_expanded_url      <list> [NA, NA, "https://twitter.com/aditrao/s…
```

Top Twitter Accounts

```
hasanIN %>%
  count(screen_name) %>%
  arrange(desc(n)) %>%
  slice(1:10)
```

```
## # A tibble: 10 x 2
##    screen_name         n
##    <chr>           <int>
##  1 IndiaAtWar         15
##  2 ViratPhoenix       14
##  3 PerzonalOpinion    12
##  4 Sreevenkat13        9
##  5 ABhadikar           7
##  6 dankchikidang       7
##  7 RollyKumari         7
##  8 swarnim_adhyaay     7
##  9 thanosisthehero     7
## 10 vedant23440716      7
```

Tweet Client Source

```r
# Tweet Client Source
hasanIN %>%
  count(source) %>%
  arrange(desc(n))
```

```
## # A tibble: 11 x 2
##    source                   n
##    <chr>                <int>
##  1 Twitter for Android    781
##  2 Twitter for iPhone     534
##  3 Twitter Web Client     244
##  4 Twitter Web App        195
##  5 Twitter for iPad        18
##  6 Mobile Web (M2)         12
##  7 TweetDeck                7
##  8 Flamingo for Android     4
##  9 Tweetbot for iOS         4
## 10 Buffer                   2
## 11 Hootsuite Inc.           2
```

Top Hashtags

```r
# Top 20 Hashtags
hasanIN %>%
  unnest(hashtags) %>%
  count(hashtags = tolower(hashtags)) %>%
  arrange(desc(n)) %>%
  mutate(hashtags = fct_reorder(hashtags,-n, .desc = TRUE)) %>%
  drop_na() %>%
  slice(1:20) %>%
  ggplot() + geom_bar(aes(hashtags,n), stat = "identity", fill = "#00
  coord_flip() +
  ggplot2::theme_minimal()+
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.text.y = element_text(face = c('bold'),
                                   size = 14,
                                   color = "#000080")) +
  labs(title = "Top 20 Hashtags about Patriot Act's Indian Election 
       subtitle = "Comdey Show by Hasan Mihnaj & Netflix",
       caption = "Data Source: Tweets mentioning `@hasanminhaj india
       y = "Count of Tweets",
       x = "Hashtags") -> top20_plot
```

The Graphics - That doesn't look interesting

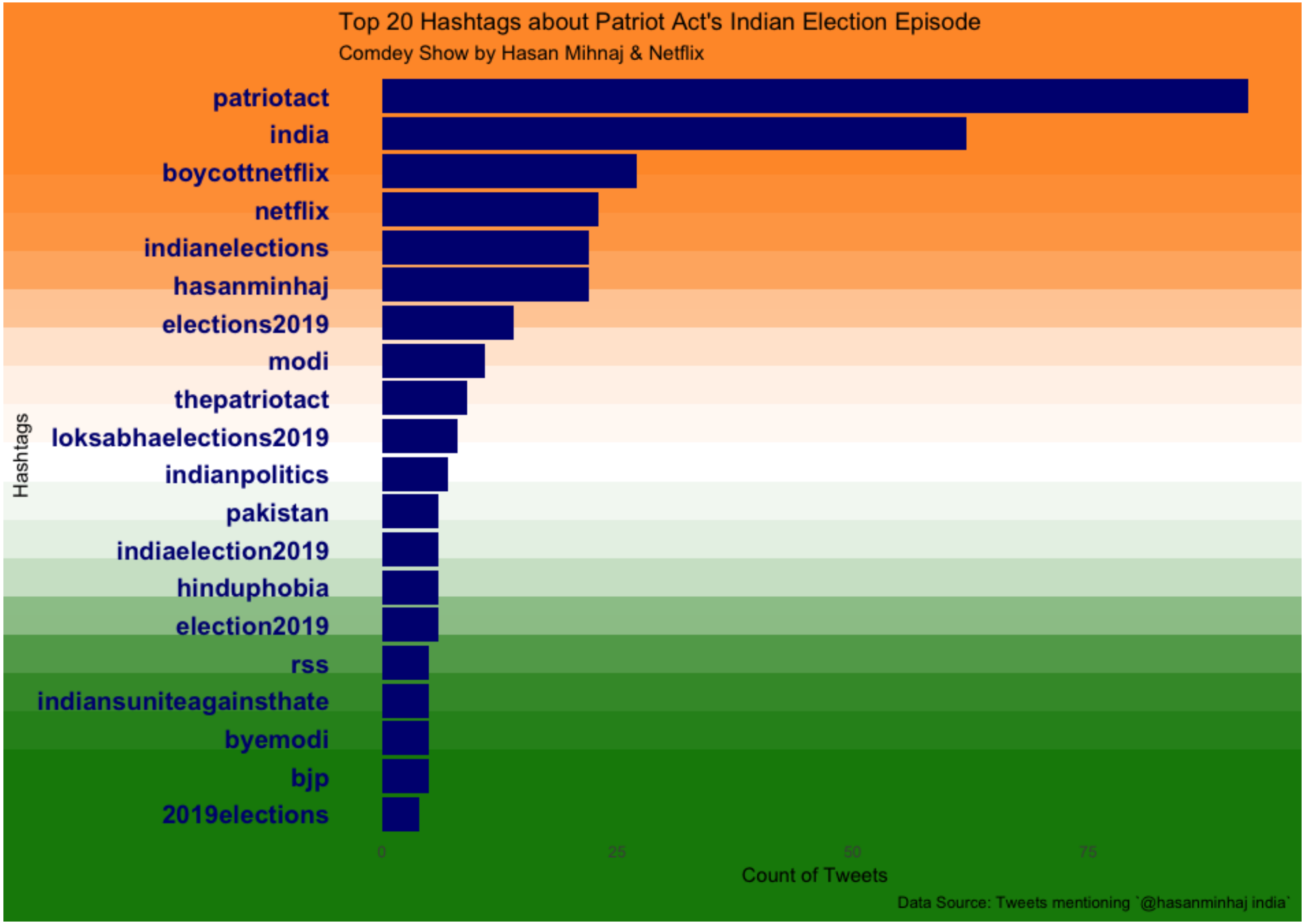
```
top20_plot
```

Themed Graphics

```
# based on this SO answer: https://stackoverflow.com/a/39632532
# Indian Tricolor Gradient Background
# Src: https://www.schemecolor.com/indian-flag-colors.php

indflag <- c("#FF9933", "#FFFFFF", "#138808")
g <- rasterGrob(indflag, width = unit(1, "npc"), height = unit(1, "np
grid.newpage()
grid.draw(g)
print(top20_plot, newpage = FALSE)
```

Themed Graphics



Top 20 Hashtags about Patriot Act's Indian Election Episode

Comdey Show by Hasan Mihnaj & Netflix

Data Source: Tweets mentioning `@hasanminhaj india`

Topic Extraction

Bit of cleaning

```r
# Cleaning

#based on: https://stackoverflow.com/questions/51947268/remove-hashta

hasanIN$text_nohashtag <- stringi::stri_replace_all_regex(hasanIN$tex
```

NLP in Action

Language Model

```
#model <- udpipe_download_model(language = "english")
udmodel_english <- udpipe_load_model(file = 'english-ewt-ud-2.3-18111
```

Annotation & Transformation

```
s <- udpipe_annotate(udmodel_english, hasanIN$text_nohashtag)

x <- data.frame(s)
```

Topic (Keyword) Extraction

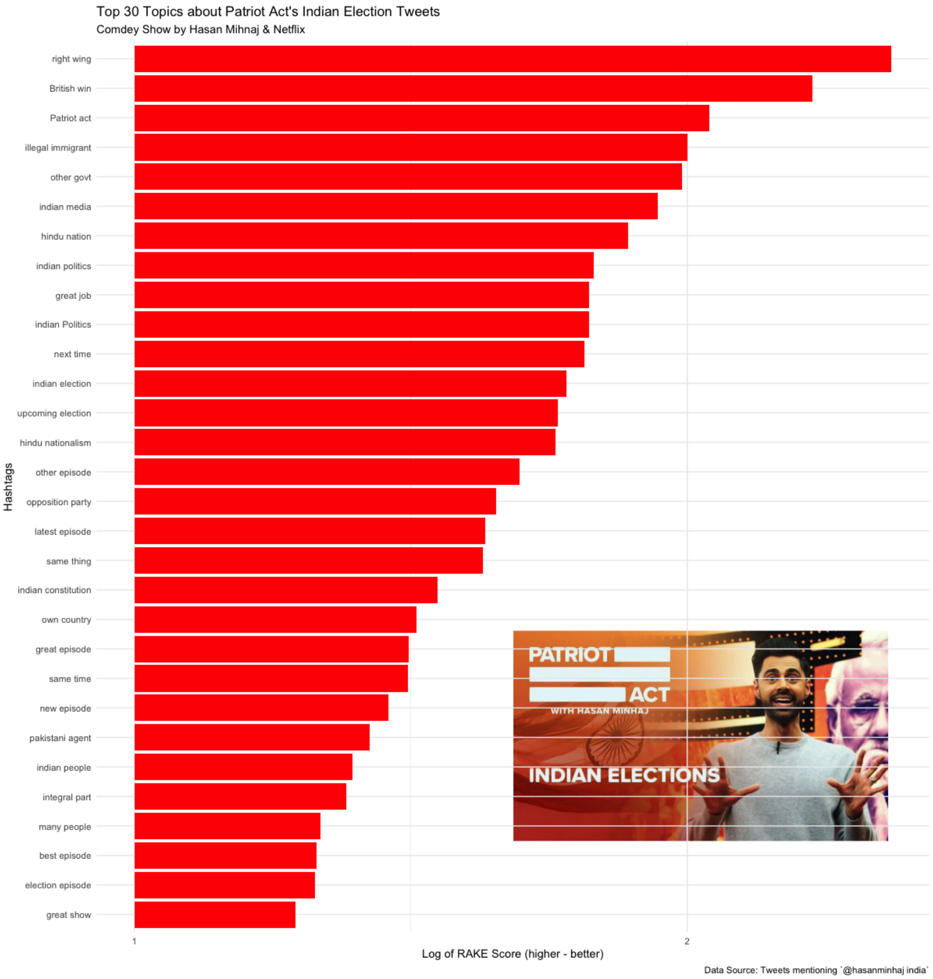
```
## Using RAKE
stats <- keywords_rake(x = x, term = "lemma", group = "doc_id",
                       relevant = x$upos %in% c("NOUN", "ADJ"))
```

Themed Graphics

```
stats %>%
  filter(freq >= 5) %>%
  arrange(desc(rake)) %>%
  slice(1:30) %>%
  mutate(keyword = fct_reorder(keyword,rake)) %>%
  ggplot() + geom_bar(aes(keyword,rake), stat = "identity", fill = "
  scale_y_log10() +
  coord_flip() +
  theme_minimal() +
  labs(title = "Top 30 Topics about Patriot Act's Indian Election Twe
       subtitle = "Comdey Show by Hasan Mihnaj & Netflix",
       caption = "Data Source: Tweets mentioning `@hasanminhaj india
       y = "Log of RAKE Score (higher - better)",
       x = "Hashtags") -> topics

ggdraw() +
  draw_image("https://st1.latestly.com/wp-content/uploads/2019/03/03
             x = 0.25, y = -0.25,
             scale = 0.4) +
  draw_plot(topics)
```

Themed Graphics

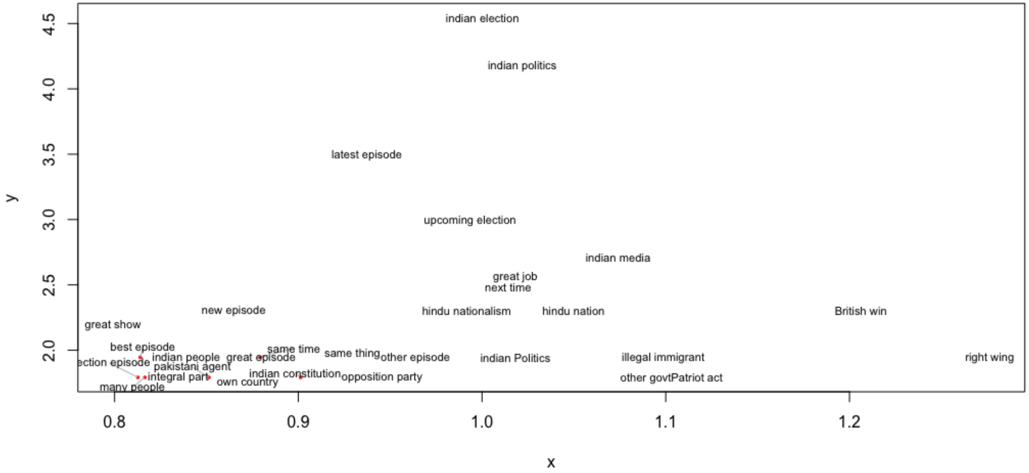


Top 30 Topics about Patriot Act's Indian Election Tweets

Comdey Show by Hasan Mihnaj & Netflix

Data Source: Tweets mentioning `@hasanminhaj india`

You can do much more!



Thanks!

Slides created via the R package **xaringan**.

The chakra comes from remark.js, **knitr**, and R Markdown.

Bibliography

```r
citation('xaringan')
```

```
## Warning in citation("xaringan"): no date field in DESCRIPTION file of
## package 'xaringan'

## Warning in citation("xaringan"): could not determine year for 'xaringan'
## from package DESCRIPTION file

## 
## To cite package 'xaringan' in publications use:
## 
##   Yihui Xie (NA). xaringan: Presentation Ninja. R package version
##   0.8.6. https://github.com/yihui/xaringan
## 
## A BibTeX entry for LaTeX users is
## 
##   @Manual{,
##     title = {xaringan: Presentation Ninja},
##     author = {Yihui Xie},
##     note = {R package version 0.8.6},
##     url = {https://github.com/yihui/xaringan},
##   }
```

Bibliography

```r
citation('udpipe')
```

```
##
## To cite package 'udpipe' in publications use:
##
##   Jan Wijffels (2019). udpipe: Tokenization, Parts of Speech
##   Tagging, Lemmatization and Dependency Parsing with the 'UDPipe'
##   'NLP' Toolkit. R package version 0.8.1.
##   https://CRAN.R-project.org/package=udpipe
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {udpipe: Tokenization, Parts of Speech Tagging, Lemmatization
## Dependency Parsing with the 'UDPipe' 'NLP' Toolkit},
##     author = {Jan Wijffels},
##     year = {2019},
##     note = {R package version 0.8.1},
##     url = {https://CRAN.R-project.org/package=udpipe},
##   }
```

Bibliography

```r
citation('tidyverse')
```

```
##
## To cite package 'tidyverse' in publications use:
##
##   Hadley Wickham (2017). tidyverse: Easily Install and Load the
##   'Tidyverse'. R package version 1.2.1.
##   https://CRAN.R-project.org/package=tidyverse
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {tidyverse: Easily Install and Load the 'Tidyverse'},
##     author = {Hadley Wickham},
##     year = {2017},
##     note = {R package version 1.2.1},
##     url = {https://CRAN.R-project.org/package=tidyverse},
##   }
```

THE END