

Sentiment Analysis with Twitter Data during the COVID-19 Pandemic

Thao Tran

2024-03-15

Source code

I. Data preparation

The Twitter dataset is available on the gesis website (<https://data.gesis.org/tweetscov19/#dataset>) and it is splitted into 3 subsets: Part 1 (Oct 2019 - April 2020), Part 2 (May 2020 - May 2020), Part 3 (June 2020 - Dec 2020).

Alternatively, the data can be accessed on the sync and share platform: <https://syncandshare.lrz.de/getlink/fiCvXkpC939CuulxEQMbmp/>

Note: the data is not provided in the GitHub repository due to its size.

```
# libraries
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(lubridate)
library(changepoint)
library(stringr) # for grepl

# read in data sets
# Note: this takes up some time (>30 min)

##### Part 1 (Oct 2019 - April 2020) #####
df_1 <- readr::read_tsv("/Users/thaotran/Downloads/Downloads/CSS/TweetsCOV19_1.tsv",
                        col_names = c("TweetID", "Username", "Timestamp", "Followers", "Friends",
                                      "Retweets", "Favorites", "Entities", "Sentiment",
                                      "Mentions", "Hashtags", "URLs"))

# seperate sentiment score
split_data <- strsplit(as.character(df_1$Sentiment), " ")

df_1$Sent_pos <- sapply(split_data, `[`, 1)
df_1$Sent_neg <- sapply(split_data, `[`, 2)

df_1$Sent_pos <- as.integer(df_1$Sent_pos)
df_1$Sent_neg <- as.integer(df_1$Sent_neg)

## format date
df_1$Date <- format(strptime(df_1$Timestamp,
                             format="%a %b %d %H:%M:%S +0000 %Y",
```

```

        tz="UTC"), "%Y-%m-%d")

##### Part 2 (May 2020 - May 2020) #####
df_2 <- readr::read_tsv("/Users/thaotran/Downloads/Downloads/CSS/TweetsCOV19_2.tsv",
                        col_names = c("TweetID", "Username", "Timestamp", "Followers", "Friends",
                                      "Retweets", "Favorites", "Entities", "Sentiment",
                                      "Mentions", "Hashtags", "URLs"))

# seperate sentiment score
split_data <- strsplit(as.character(df_2$Sentiment), " ")

df_2$Sent_pos <- sapply(split_data, `[`, 1)
df_2$Sent_neg <- sapply(split_data, `[`, 2)

df_2$Sent_pos <- as.integer(df_2$Sent_pos)
df_2$Sent_neg <- as.integer(df_2$Sent_neg)

## format date
df_2$Date <- format(strptime(df_2$Timestamp,
                             format="%a %b %d %H:%M:%S +0000 %Y",
                             tz="UTC"), "%Y-%m-%d")

##### Part 3 (June 2020 - Dec 2020) #####
df_3 <- readr::read_tsv("/Users/thaotran/Downloads/Downloads/CSS/TweetsCOV19_3.tsv",
                        col_names = c("TweetID", "Username", "Timestamp", "Followers", "Friends",
                                      "Retweets", "Favorites", "Entities", "Sentiment",
                                      "Mentions", "Hashtags", "URLs"))

# seperate sentiment score
split_data <- strsplit(as.character(df_3$Sentiment), " ")

df_3$Sent_pos <- sapply(split_data, `[`, 1)
df_3$Sent_neg <- sapply(split_data, `[`, 2)

df_3$Sent_pos <- as.integer(df_3$Sent_pos)
df_3$Sent_neg <- as.integer(df_3$Sent_neg)

## format date
df_3$Date <- format(strptime(df_3$Timestamp,
                             format="%a %b %d %H:%M:%S +0000 %Y",
                             tz="UTC"),
                    "%Y-%m-%d")

# bind whole data set
df_s <- bind_rows(df_1, df_2, df_3)

```

II. Analysis

1) Number of tweets

```

count_per_day <- df_s %>%
  group_by(Date) %>%

```

```

    summarise(Count = n())

count_per_day$Date <-as.Date(count_per_day$Date)

# plot
no <- ggplot(data=count_per_day, aes(x=Date, y=Count)) +
  geom_bar(stat="identity") +
  labs(title = "", x = "Date", y = "Total Number of Tweets per Day") +
  geom_vline(xintercept = as.Date("2020-04-01"), linetype = "dashed", color = "steelblue",
    size = 0.8) +
  annotate("text", x = as.Date("2020-04-20"), y = 120000,
    label = "April 2020:\nmmost countries under lockdown", color = "steelblue",
    hjust = 0, size = 3.5) +
  scale_x_date(breaks = seq(min(count_per_day$Date)+1, max(count_per_day$Date),
    by = "2 months"), date_labels = "%Y-%m") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
no

```

2) Sentiment over time

```

# Mean per day
mean_sent_per_day <- df_s %>%
  group_by(Date) %>%
  summarise(
    Mean_Sent_Pos = mean(Sent_pos, na.rm = TRUE),
    Mean_Sent_Neg = mean(Sent_neg, na.rm = TRUE)
  )
mean_sent_per_day$Date <-as.Date(mean_sent_per_day$Date)

# plot: positive sentiment
pos <- ggplot(mean_sent_per_day, aes(Date)) +
  geom_line(aes(y = Mean_Sent_Pos)) +
  geom_vline(xintercept = as.Date("2020-04-01"), linetype = "dashed", color = "steelblue",
    size = 0.8) +
  ylim(1.45, 1.8) +
  labs(title = "Positive Sentiment", x = "Date", y = "Mean per Day") +
  annotate("text", x = as.Date("2020-04-10"), y = 1.77,
    label = "April 2020:\nmmost countries under lockdown", color = "steelblue",
    hjust = 0, size = 3.5) +
  scale_x_date(breaks = seq(min(mean_sent_per_day$Date)+1, max(mean_sent_per_day$Date),
    by = "2 months"), date_labels = "%Y-%m") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
pos

# plot: negative sentiment
neg <- ggplot(mean_sent_per_day, aes(Date)) +
  geom_line(aes(y = Mean_Sent_Neg)) +
  geom_vline(xintercept = as.Date("2020-04-01"), linetype = "dashed", color = "steelblue",
    size = 0.8) +
  ylim(-1.8, -1.45) +
  labs(title = "Negative Sentiment", x = "Date", y = "Mean per Day") +
  annotate("text", x = as.Date("2020-04-10"), y = -1.48,

```

```

        label = "April 2020:\nmost countries under lockdown", color = "steelblue",
        hjust = 0, size = 3.5) +
scale_x_date(breaks = seq(min(mean_sent_per_day$Date)+1, max(mean_sent_per_day$Date),
                          by = "2 months"), date_labels = "%Y-%m") +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
neg

#### TREND - Sentiment over time - POS ####
trend_sent_pos <- cpt.meanvar(mean_sent_per_day$Mean_Sent_Pos, method="BinSeg", Q=3)
trend_sent_pos
plot(trend_sent_pos)

as.Date("2019-10-01") + 115 # "2020-01-24"
as.Date("2019-10-01") + 187 # "2020-04-05"
as.Date("2019-10-01") + 306 # "2020-08-02"
trend_sent_pos@param.est

trend_pos <- ggplot(mean_sent_per_day, aes(Date)) +
  geom_line(aes(y = Mean_Sent_Pos)) +
  geom_vline(xintercept = as.Date("2020-04-01"), linetype = "dashed", color = "steelblue",
            size = 0.8) + ylim(1.45, 1.8) +
  labs(title = "Positive Sentiment", x = "Date", y = "Mean per Day") +
  geom_segment(aes(x = as.Date("2019-09-30"), xend = as.Date("2020-01-24"),
                  y = 1.631671, yend = 1.631671),
              linetype = "solid", color = "indianred", size = 1) +
  geom_segment(aes(x = as.Date("2020-01-24"), xend = as.Date("2020-04-05"),
                  y = 1.572026, yend = 1.572026),
              linetype = "solid", color = "indianred", size = 1) +
  geom_segment(aes(x = as.Date("2020-04-05"), xend = as.Date("2020-08-02"),
                  y = 1.581613, yend = 1.581613),
              linetype = "solid", color = "indianred", size = 1) +
  geom_segment(aes(x = as.Date("2020-08-02"), xend = as.Date("2020-12-31"),
                  y = 1.596436, yend = 1.596436),
              linetype = "solid", color = "indianred", size = 1) +
  annotate("text", x = as.Date("2020-04-10"), y = 1.77,
          label = "April 2020:\nmost countries under lockdown", color = "steelblue",
          hjust = 0, size = 3.5) +
scale_x_date(breaks = seq(min(mean_sent_per_day$Date)+1, max(mean_sent_per_day$Date),
                          by = "2 months"), date_labels = "%Y-%m") +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
trend_pos

##### TREND - Sentiment over time - NEG #####
trend_sent_neg <- cpt.meanvar(mean_sent_per_day$Mean_Sent_Neg, method="BinSeg", Q=4)
trend_sent_neg
plot(trend_sent_neg)

as.Date("2019-10-01") + 66 # "2019-12-06"
as.Date("2019-10-01") + 164 # "2020-03-13"
as.Date("2019-10-01") + 240 # "2020-05-28"
as.Date("2019-10-01") + 252 # "2020-06-09"
trend_sent_neg@param.est

```

```

trend_neg <- ggplot(mean_sent_per_day, aes(Date)) +
  geom_line(aes(y = Mean_Sent_Neg)) +
  geom_vline(xintercept = as.Date("2020-04-01"), linetype = "dashed", color = "steelblue",
    size = 0.8) +
  ylim(-1.8, -1.45) +
  labs(title = "Negative Sentiment", x = "Date", y = "Mean per Day") +
  annotate("text", x = as.Date("2020-04-10"), y = -1.48,
    label = "April 2020:\nmmost countries under lockdown", color = "steelblue",
    hjust = 0, size = 3.5) +
  geom_segment(aes(x = as.Date("2019-09-30"), xend = as.Date("2019-12-06"),
    y = -1.617850, yend = -1.617850),
    linetype = "solid", color = "indianred", size = 1) +
  geom_segment(aes(x = as.Date("2019-12-06"), xend = as.Date("2020-03-13"),
    y = -1.635428, yend = -1.635428),
    linetype = "solid", color = "indianred", size = 1) +
  geom_segment(aes(x = as.Date("2020-03-13"), xend = as.Date("2020-05-28"),
    y = -1.595235, yend = -1.595235),
    linetype = "solid", color = "indianred", size = 1) +
  geom_segment(aes(x = as.Date("2020-05-28"), xend = as.Date("2020-06-09"),
    y = -1.728316, yend = -1.728316),
    linetype = "solid", color = "indianred", size = 1) +
  geom_segment(aes(x = as.Date("2020-06-09"), xend = as.Date("2020-12-31"),
    y = -1.631783, yend = -1.631783),
    linetype = "solid", color = "indianred", size = 1) +
  scale_x_date(breaks = seq(min(mean_sent_per_day$Date)+1, max(mean_sent_per_day$Date),
    by = "2 months"), date_labels = "%Y-%m") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
trend_neg

```

3) Specific words

a) stem “depress”

```

# Using grepl to find 'depress'
df_s$contains_depress <- as.numeric(grepl("depress", df_s$Entities, ignore.case = TRUE))

# calculate occurrences (total and share)
result_depress <- df_s %>%
  group_by(Date) %>%
  summarise(
    count_depress = sum(contains_depress),
    total_count = n(),
    proportion_depress = count_depress / total_count
  )
result_depress$Date <- as.Date(result_depress$Date)

depress1 <- ggplot(data=result_depress, aes(x=Date, y=proportion_depress)) +
  geom_bar(stat="identity") +
  labs(title = "", x = "Date", y = "Share of Tweets containing the stem 'depress' ") +
  geom_vline(xintercept = as.Date("2020-04-01"), linetype = "dashed", color = "steelblue",
    size = 0.8) +

```

```

    annotate("text", x = as.Date("2020-04-10"), y = 0.00054,
             label = "April 2020:\nmmost countries under lockdown", color = "steelblue",
             hjust = 0, size = 3.5) +
    scale_x_date(breaks = seq(min(result_depress$Date)+1, max(result_depress$Date),
                             by = "2 months"), date_labels = "%Y-%m") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
depress1

#### TREND - DEPRESS ####
trend_depress <- cpt.meanvar(result_depress$proportion_depress, method="BinSeg")
trend_depress
plot(trend_depress)

as.Date("2019-10-01") + 11 # "2019-10-12" --> mental health day
as.Date("2019-10-01") + 192 # "2020-04-10"
as.Date("2019-10-01") + 259 # "2020-06-16"

trend_depress@param.est

depress2 <- ggplot(data=result_depress, aes(x=Date, y=proportion_depress)) +
  geom_bar(stat="identity") +
  labs(title = "", x = "Date", y = "Share of Tweets containing 'depress'") +
  geom_vline(xintercept = as.Date("2020-04-01"), linetype = "dashed", color = "steelblue",
             size = 0.8) +
  annotate("text", x = as.Date("2020-04-10"), y = 0.00053,
           label = "April 2020:\nmmost countries under lockdown", color = "steelblue",
           hjust = 0, size = 3.5) +
  geom_segment(aes(x = as.Date("2019-09-30"), xend = as.Date("2019-10-12"),
                   y = 0.0002495771, yend = 0.0002495771),
               linetype = "solid", color = "indianred", size = 1) +
  geom_segment(aes(x = as.Date("2019-10-12"), xend = as.Date("2020-04-10"),
                   y = 0.0001348627, yend = 0.0001348627),
               linetype = "solid", color = "indianred", size = 1) +
  geom_segment(aes(x = as.Date("2020-04-10"), xend = as.Date("2020-06-16"),
                   y = 0.0002289855, yend = 0.0002289855),
               linetype = "solid", color = "indianred", size = 1) +
  geom_segment(aes(x = as.Date("2020-06-16"), xend = as.Date("2020-12-31"),
                   y = 0.0001431119, yend = 0.0001431119),
               linetype = "solid", color = "indianred", size = 1) +
  scale_x_date(breaks = seq(min(result_depress$Date)+1, max(result_depress$Date),
                             by = "2 months"), date_labels = "%Y-%m") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
depress2

```

b) “love”

```

# Using grepl to find 'love'
df_s$contains_love <- as.numeric(grepl("(^[^a-zA-Z])love", df_s$Entities, ignore.case = TRUE))

# calculate occurrences (total and share)
result_love <- df_s %>%
  group_by(Date) %>%

```

```

    summarise(
      count_love = sum(contains_love),
      total_count = n(),
      proportion_love = count_love / total_count
    )
result_love$Date <-as.Date(result_love$Date)

love1 <- ggplot(data=result_love, aes(x=Date, y=proportion_love)) +
  geom_bar(stat="identity") +
  labs(title = "", x = "Date", y = "Share of Tweets containing 'love'") +
  geom_vline(xintercept = as.Date("2020-04-01"), linetype = "dashed", color = "steelblue",
    size = 0.8) +
  annotate("text", x = as.Date("2020-04-10"), y = 0.007,
    label = "April 2020:\nmmost countries under lockdown", color = "steelblue",
    hjust = 0, size = 3.5) +
  scale_x_date(breaks = seq(min(result_love$Date)+1, max(result_love$Date),
    by = "2 months"), date_labels = "%Y-%m") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
love1

#### TREND - LOVE ####
trend_love <- cpt.meanvar(result_love$proportion_love, method="BinSeg")
trend_love
plot(trend_love)

as.Date("2019-10-01") + 128 # "2020-02-06" --> increase bc of valentine's day ?
as.Date("2019-10-01") + 147 # "2020-02-25"
as.Date("2019-10-01") + 334 # "2020-08-30"
trend_love@param.est

love2 <- ggplot(data=result_love, aes(x=Date, y=proportion_love)) +
  geom_bar(stat="identity") +
  labs(title = "", x = "Date", y = "Share of Tweets containing 'love'") +
  geom_vline(xintercept = as.Date("2020-04-01"), linetype = "dashed", color = "steelblue",
    size = 0.8) +
  annotate("text", x = as.Date("2020-04-10"), y = 0.007,
    label = "April 2020:\nmmost countries under lockdown", color = "steelblue",
    hjust = 0, size = 3.5) +
  geom_segment(aes(x = as.Date("2019-09-30"), xend = as.Date("2020-02-06"),
    y = 0.003483227, yend = 0.003483227),
    linetype = "solid", color = "indianred", size = 1) +
  geom_segment(aes(x = as.Date("2020-02-06"), xend = as.Date("2020-02-25"),
    y = 0.003689664, yend = 0.003689664),
    linetype = "solid", color = "indianred", size = 1) +
  geom_segment(aes(x = as.Date("2020-02-25"), xend = as.Date("2020-08-30"),
    y = 0.002345468, yend = 0.002345468),
    linetype = "solid", color = "indianred", size = 1) +
  geom_segment(aes(x = as.Date("2020-08-30"), xend = as.Date("2020-12-31"),
    y = 0.002656875, yend = 0.002656875),
    linetype = "solid", color = "indianred", size = 1) +
  scale_x_date(breaks = seq(min(result_love$Date)+1, max(result_love$Date),
    by = "2 months"), date_labels = "%Y-%m") +

```

```
theme(axis.text.x = element_text(angle = 45, hjust = 1))
love2
```

c) “hate”

```
# Using grepl to find 'hate'
df_s$contains_hate <- as.numeric(grepl("\\b(hate[~a-zA-Z]|hate$)", df_s$Entities,
                                     ignore.case = TRUE))

# calculate occurrences (total and share)
result_hate <- df_s %>%
  group_by(Date) %>%
  summarise(
    count_hate = sum(contains_hate),
    total_count = n(),
    proportion_hate = count_hate / total_count
  )
result_hate$Date <- as.Date(result_hate$Date)

hate1 <- ggplot(data=result_hate, aes(x=Date, y=proportion_hate)) +
  geom_bar(stat="identity") +
  ylim(0, 0.00075) +
  labs(title = "", x = "Date", y = "Share of Tweets") +
  geom_vline(xintercept = as.Date("2020-04-01"), linetype = "dashed", color = "steelblue",
            size = 0.8) +
  annotate("text", x = as.Date("2020-04-10"), y = 0.0016,
          label = "April 2020:\nmmost countries under lockdown", color = "steelblue",
          hjust = 0, size = 3.5) +
  scale_x_date(breaks = seq(min(result_hate$Date)+1, max(result_hate$Date),
                            by = "2 months"), date_labels = "%Y-%m") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
hate1

#### TREND - HATE ####
trend_hate <- cpt.meanvar(result_hate$proportion_hate, method="BinSeg")
trend_hate
plot(trend_hate)

as.Date("2019-10-01") + 287 # "2020-07-14"
as.Date("2019-10-01") + 406 # "2020-11-10"
trend_hate@param.est

hate2 <- ggplot(data=result_hate, aes(x=Date, y=proportion_hate)) +
  geom_bar(stat="identity") +
  labs(title = "", x = "Date", y = "Share of Tweets containing 'hate'") +
  geom_vline(xintercept = as.Date("2020-04-01"), linetype = "dashed", color = "steelblue",
            size = 0.8) +
  annotate("text", x = as.Date("2020-04-10"), y = 0.0016,
          label = "April 2020:\nmmost countries under lockdown", color = "steelblue",
          hjust = 0, size = 3.5) +
  geom_segment(aes(x = as.Date("2019-09-30"), xend = as.Date("2020-07-14"),
                  y = 0.0003738597, yend = 0.0003738597),
```



```

        linetype = "solid", color = "indianred", size = 1) +
geom_segment(aes(x = as.Date("2020-07-14"), xend = as.Date("2020-11-10"),
                y = 0.0003465534, yend = 0.0003465534),
        linetype = "solid", color = "indianred", size = 1) +
geom_segment(aes(x = as.Date("2020-11-10"), xend = as.Date("2020-12-31"),
                y = 0.0002550559, yend = 0.0002550559),
        linetype = "solid", color = "indianred", size = 1) +
scale_x_date(breaks = seq(min(result_hate$Date)+1, max(result_hate$Date),
                        by = "2 months"), date_labels = "%Y-%m") +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
hate2

```

4) Popular tweets

Tweets with 200 retweets or more.

```

retweets_200 <- df_s[df_s$Retweets >= 200, ] # 200: 627464/19969565 = 0.03142 = 3.142%

# calculate mean sentiment for each day
mean_poptweets <- retweets_200 %>%
  group_by(Date) %>%
  summarise(
    Mean_Sent_Pos = mean(Sent_pos, na.rm = TRUE),
    Mean_Sent_Neg = mean(Sent_neg, na.rm = TRUE)
  )
mean_poptweets$Date <- as.Date(mean_poptweets$Date)

pos <- ggplot(mean_poptweets, aes(Date)) +
  geom_line(aes(y = Mean_Sent_Pos)) +
  geom_vline(xintercept = as.Date("2020-04-01"), linetype = "dashed", color = "steelblue",
            size = 0.5) +
  ylim(1, 2.2) +
  labs(title = "Positive Sentiment", x = "Date", y = "Mean per Day") +
  annotate("text", x = as.Date("2020-04-10"), y = 2.1,
          label = "April 2020:\nmmost countries under lockdown", color = "steelblue",
          hjust = 0, size = 3.5) +
  scale_x_date(breaks = seq(min(mean_poptweets$Date)+1, max(mean_poptweets$Date),
                        by = "2 months"), date_labels = "%Y-%m") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
pos

neg <- ggplot(mean_poptweets, aes(Date)) +
  geom_line(aes(y = Mean_Sent_Neg)) +
  geom_vline(xintercept = as.Date("2020-04-01"), linetype = "dashed", color = "steelblue",
            size = 0.5) +
  ylim(-3.2, -0.9) +
  labs(title = "Negative Sentiment", x = "Date", y = "Mean per Day") +
  annotate("text", x = as.Date("2020-04-10"), y = -1,
          label = "April 2020:\nmmost countries under lockdown", color = "steelblue",
          hjust = 0, size = 3.5) +
  scale_x_date(breaks = seq(min(mean_poptweets$Date)+1, max(mean_poptweets$Date),
                        by = "2 months"), date_labels = "%Y-%m") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

```

neg

#### positive trend ####
trend_poptweets_pos <- cpt.meanvar(mean_poptweets$Mean_Sent_Pos, method="BinSeg")
trend_poptweets_pos
plot(trend_poptweets_pos)

as.Date("2019-10-01") + 114 # "2020-01-23"
as.Date("2019-10-01") + 148 # "2020-02-26"
as.Date("2019-10-01") + 362 # "2020-09-27"
trend_poptweets_pos@param.est

trend_poptw_pos <- ggplot(mean_poptweets, aes(Date)) +
  geom_line(aes(y = Mean_Sent_Pos)) +
  geom_vline(xintercept = as.Date("2020-04-01"), linetype = "dashed", color = "steelblue",
    size = 0.5) +
  ylim(1.5, 2.2) +
  labs(title = "Positive Sentiment", x = "Date", y = "Mean per Day") +
  annotate("text", x = as.Date("2020-04-10"), y = 2.1,
    label = "April 2020:\nmmost countries under lockdown", color = "steelblue",
    hjust = 0, size = 3.5) +
  geom_segment(aes(x = as.Date("2019-09-30"), xend = as.Date("2020-01-23"),
    y = 1.636013, yend = 1.636013),
    linetype = "solid", color = "indianred", size = 1) +
  geom_segment(aes(x = as.Date("2020-01-23"), xend = as.Date("2020-02-26"),
    y = 1.598912, yend = 1.598912),
    linetype = "solid", color = "indianred", size = 1) +
  geom_segment(aes(x = as.Date("2020-02-26"), xend = as.Date("2020-09-27"),
    y = 1.568102, yend = 1.568102),
    linetype = "solid", color = "indianred", size = 1) +
  geom_segment(aes(x = as.Date("2020-09-27"), xend = as.Date("2020-12-31"),
    y = 1.595089, yend = 1.595089),
    linetype = "solid", color = "indianred", size = 1) +
  scale_x_date(breaks = seq(min(mean_poptweets$Date)+1, max(mean_poptweets$Date),
    by = "2 months"), date_labels = "%Y-%m") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
trend_poptw_pos

#### negative trend ####
trend_poptweets_neg <- cpt.meanvar(mean_poptweets$Mean_Sent_Neg, method="BinSeg", Q = 3)
trend_poptweets_neg
plot(trend_poptweets_neg)

as.Date("2019-10-01") + 151 # "2020-01-10"
as.Date("2019-10-01") + 239 # "2020-01-10"
as.Date("2019-10-01") + 281 # "2020-01-10"
trend_poptweets_neg@param.est

trend_poptw_neg <- ggplot(mean_poptweets, aes(Date)) +
  geom_line(aes(y = Mean_Sent_Neg)) +
  geom_vline(xintercept = as.Date("2020-04-01"), linetype = "dashed", color = "steelblue",
    size = 0.5) +
  ylim(-2.2, -1.5) +

```

```

labs(title = "Negative Sentiment", x = "Date", y = "Mean per Day") +
annotate("text", x = as.Date("2020-04-10"), y = -1.55,
        label = "April 2020:\nmmost countries under lockdown", color = "steelblue",
        hjust = 0, size = 3.5) +
geom_segment(aes(x = as.Date("2019-09-30"), xend = as.Date("2019-10-12"),
                y = -1.762700, yend = -1.762700),
            linetype = "solid", color = "indianred", size = 1) +
geom_segment(aes(x = as.Date("2019-10-12"), xend = as.Date("2020-04-10"),
                y = -1.724817, yend = -1.724817),
            linetype = "solid", color = "indianred", size = 1) +
geom_segment(aes(x = as.Date("2020-04-10"), xend = as.Date("2020-06-16"),
                y = -1.832427, yend = -1.832427),
            linetype = "solid", color = "indianred", size = 1) +
geom_segment(aes(x = as.Date("2020-06-16"), xend = as.Date("2020-12-31"),
                y = -1.717869, yend = -1.717869),
            linetype = "solid", color = "indianred", size = 1) +
scale_x_date(breaks = seq(min(mean_poptweets$Date)+1, max(mean_poptweets$Date),
                        by = "2 months"), date_labels = "%Y-%m") +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
trend_poptw_neg

```