# Sentiment Analysis - Ebola

*Ajinkya Sheth*

*June 13, 2019*

```
trigger <- read.csv(file="./Trigger_Other.csv", header=TRUE, sep=",", stringsAsFactors = FALSE)
```

```
concerns=trigger$t_q6
questions=trigger$t_q7
risks=trigger$t_q8
byelaws=trigger$t_q9
discussions=trigger$t_q10
capability=trigger$t_q11
```

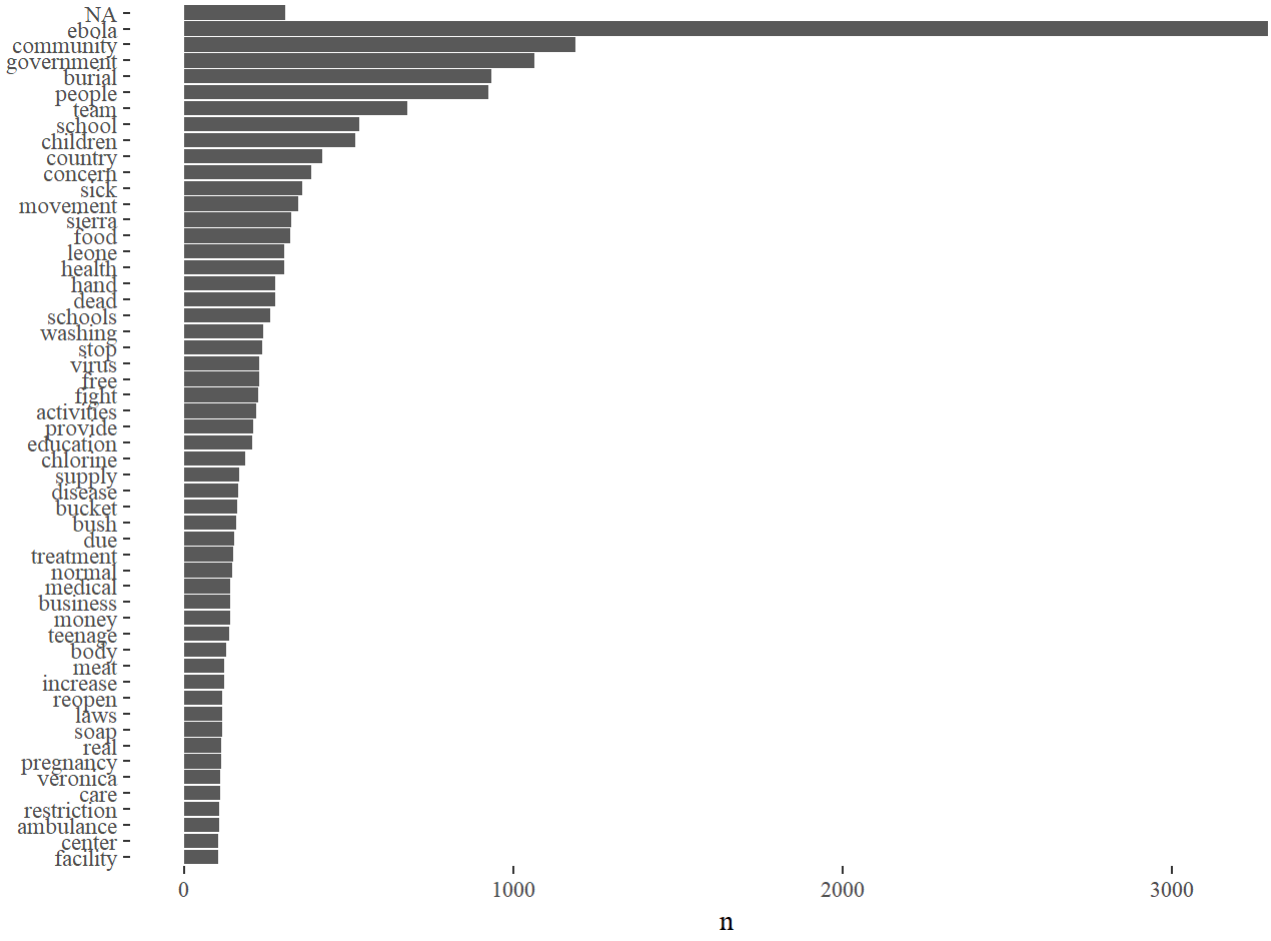
```
concerns_df <- tibble(line = 1:length(concerns),text=concerns )
tidy_corpus <- concerns_df %>%
  unnest_tokens(word, text)
data('stop_words')
tidy_corpus <- tidy_corpus %>%
  anti_join(stop_words)
```

```
## Joining, by = "word"
```

# 1 Explorations

## 1.1 Word Counts

```
tidy_corpus %>%
  count(word, sort = TRUE) %>%
  filter(n > 100) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
    geom_col() +
    xlab(NULL) +
    coord_flip() +
  theme_tufte()
```




```
tidy_corpus=tidy_corpus[tidy_corpus$word != "ebola",]
tidy_corpus=tidy_corpus[!is.na(tidy_corpus$word),]

corpus_sentiment <- tidy_corpus %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative)
```

```
## Joining, by = "word"
```

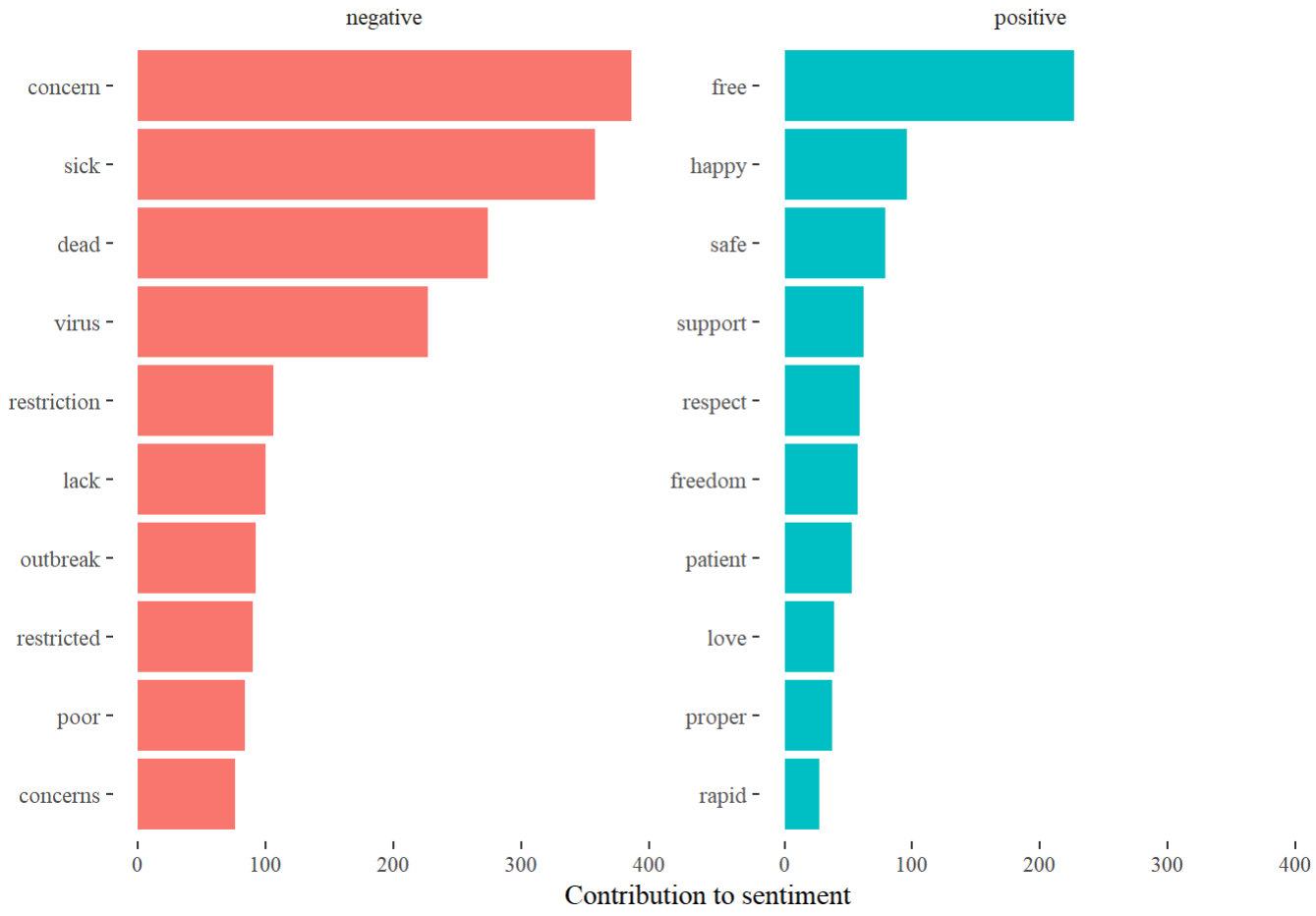
```
word_counts <- tidy_corpus %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup()
```

```
## Joining, by = "word"
```

# 1.2 Positive and Negative - Word Frequencies

```
positive_sentiment <- corpus_sentiment %>%
  filter(sentiment>0) %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup()

word_counts %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(y = "Contribution to sentiment",
       x = NULL) +
  coord_flip() +
  theme_tufte()
```

```
## Selecting by n
```



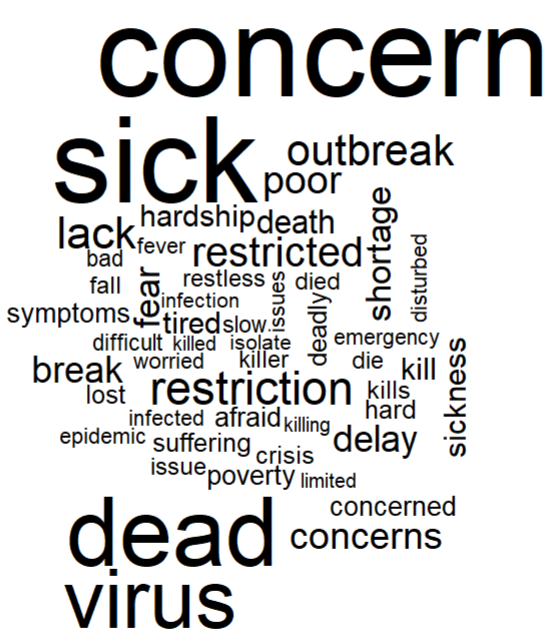

# 1.3 Positive and Negative - Word Clouds

```
layout(matrix(c(1, 2), nrow=2), heights=c(1, 6))
par(mar=rep(0, 4))
plot.new()
text(x=0.5, y=0.5, "Positive Sentiment")

word_counts %>%
  filter(sentiment=='positive') %>%
  with(wordcloud(word, n, max.words = 50))
```

Positive Sentiment


```
layout(matrix(c(1, 2), nrow=2), heights=c(1, 6))
par(mar=rep(0, 4))
plot.new()
text(x=0.5, y=0.5, "Negative Sentiment")
word_counts %>%
  filter(sentiment=='negative') %>%
  with(wordcloud(word, n, max.words = 50))
```

Negative Sentiment




## 1.4 Sentimentr tests

sentimentr tests

```
sentences_list <- c("I like ice-cream",
                    "I do not like ice-cream",
                    "I hate ice-cream",
                    "I hardly like ice-cream",
                    "Do you think I like ice-cream?")

sentences_list %>%
  get_sentences() %>%
  sentiment
```

```
##    element_id sentence_id word_count   sentiment
## 1:          1           1          4  0.25000000
## 2:          2           1          6 -0.20412415
## 3:          3           1          4 -0.37500000
## 4:          4           1          5  0.04472136
## 5:          5           1          7  0.18898224
```

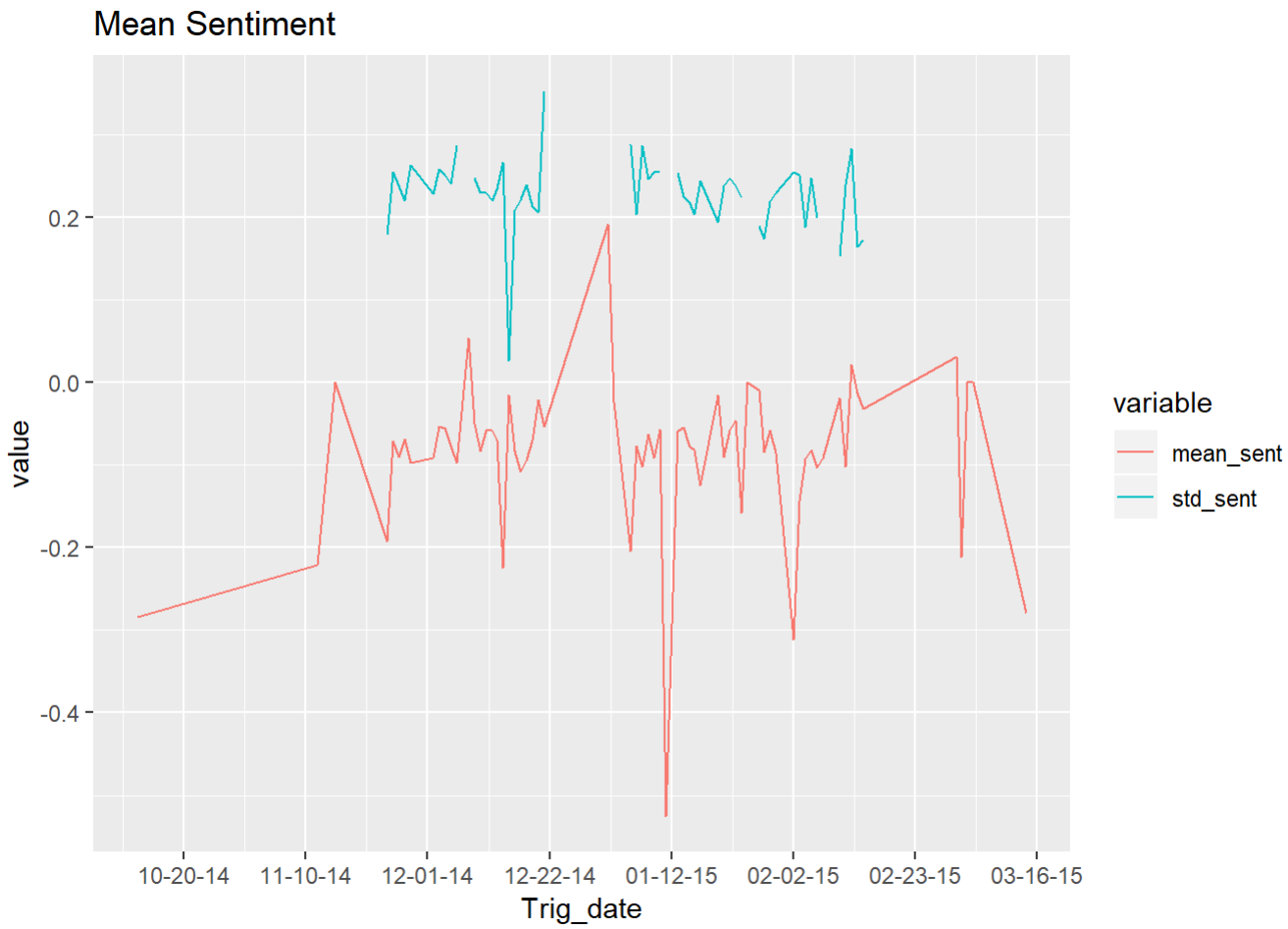
Notes for Sentiment by time: - Two fields: mean sentiment as a function of a day - Approx. Positive slope - Two extremely negative peaks, indicate sentiment about concerns over entire country fell - Terrible news can be a confounding factor

# 2 Sentiment by time

## 2.1 Mean Sentiment as a function of day

```
A <-  (trigger %>%
  dplyr::select(Trig_date,District,Chiefdom,t_q6))

B <-  (trigger$t_q6 %>%
         get_sentences() %>%
  sentiment_by(by=NULL))

sentiment_concerns <- cbind(A,B)

sentiment_concerns$Trig_date <- as.Date(sentiment_concerns$Trig_date, "%m/%d/%Y")

concerns_by_day <- sentiment_concerns %>%
  dplyr::select(Trig_date,ave_sentiment)

concerns_by_day <- concerns_by_day %>%
  group_by(Trig_date) %>%
  summarise(mean_sent = mean(ave_sentiment, na.rm = TRUE), std_sent = sd(ave_sentiment, na.rm =
TRUE))

require(ggplot2)
ggplot( data = concerns_by_day, aes( x=Trig_date, y=value, color=variable)) +
  geom_line(aes(y=mean_sent, color = "mean_sent")) +
  geom_line(aes(y=std_sent, color = "std_sent")) +
  ggtitle('Mean Sentiment') +
  scale_x_date(labels = date_format("%m-%d-%y"), breaks = "3 week")
```

```
## Warning: Removed 1 rows containing missing values (geom_path).
```



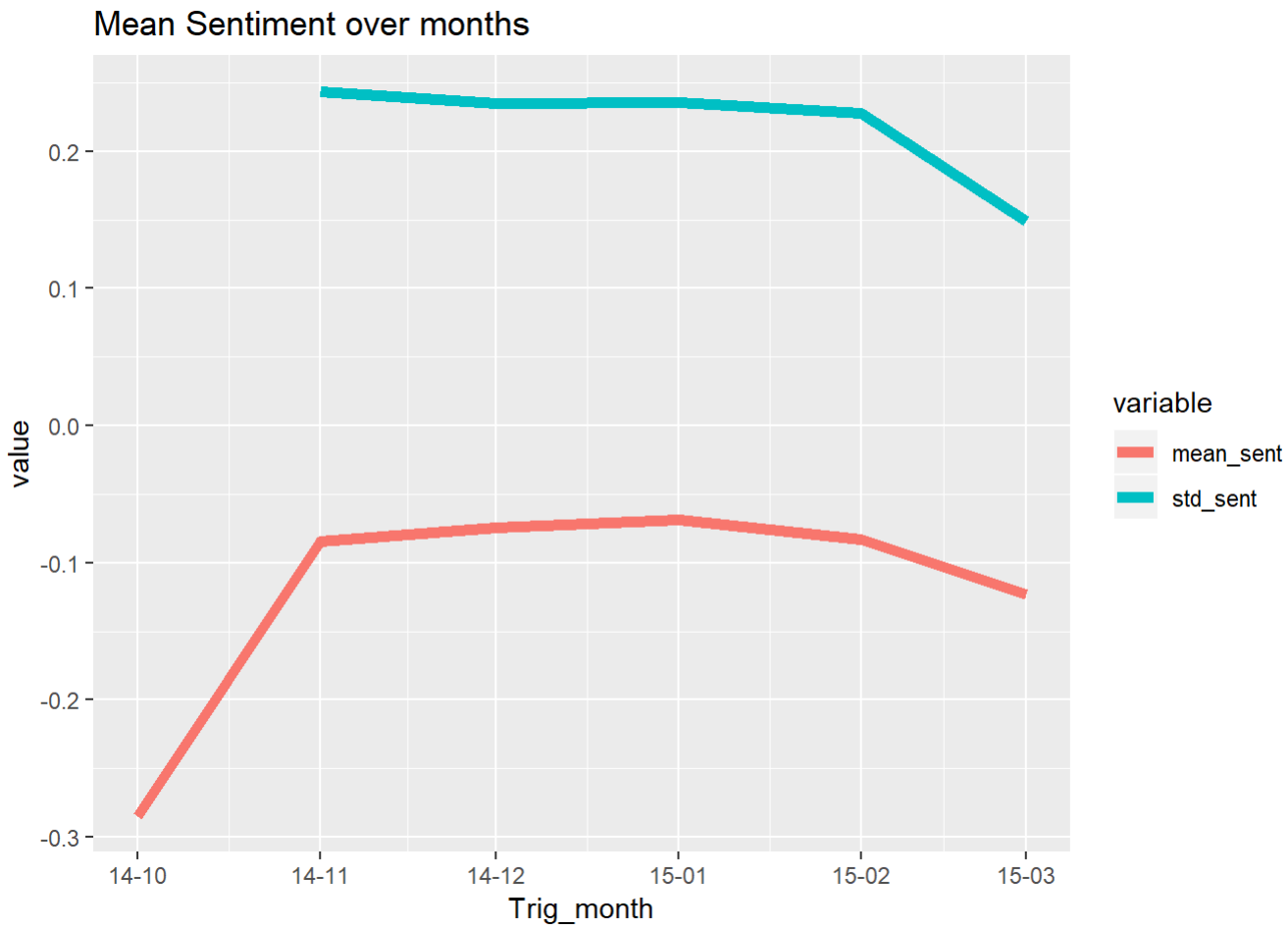

```
#geom_point()

concerns_by_day
```

```
## # A tibble: 67 x 3
##    Trig_date  mean_sent std_sent
##    <date>         <dbl>    <dbl>
##  1 2014-10-12   -0.285   NaN
##  2 2014-11-12   -0.221     0.265
##  3 2014-11-15    0       NaN
##  4 2014-11-24   -0.192     0.179
##  5 2014-11-25   -0.0712    0.255
##  6 2014-11-26   -0.0899    0.238
##  7 2014-11-27   -0.0691    0.220
##  8 2014-11-28   -0.0978    0.263
##  9 2014-12-02   -0.0912    0.228
## 10 2014-12-03   -0.0530    0.258
## # ... with 57 more rows
```

## 2.2 Mean Sentiment as a function of month

```
A <-  (trigger %>%
  dplyr::select(Trig_date,District,Chiefdom,t_q6))

B <-  (trigger$t_q6 %>%
         get_sentences() %>%
  sentiment_by(by=NULL))

sentiment_concerns_cuts <- cbind(A,B)

sentiment_concerns_cuts$Trig_date <- as.Date(sentiment_concerns_cuts$Trig_date, "%m/%d/%Y")
sentiment_concerns_cuts$Trig_month <- as.Date(cut(sentiment_concerns_cuts$Trig_date,breaks = "mo
nth"))
sentiment_concerns_cuts$Trig_week <- as.Date(cut(sentiment_concerns_cuts$Trig_date,breaks = "wee
k",start.on.monday = FALSE))

concerns_by_week <- sentiment_concerns_cuts %>%
  dplyr::select(Trig_week,ave_sentiment)

concerns_by_week <- concerns_by_week %>%
  group_by(Trig_week) %>%
  summarise(mean_sent = mean(ave_sentiment, na.rm = TRUE), std_sent = sd(ave_sentiment, na.rm =
TRUE))

concerns_by_month <- sentiment_concerns_cuts %>%
  dplyr::select(Trig_month, ave_sentiment)

concerns_by_month <- concerns_by_month %>%
  group_by(Trig_month) %>%
  summarise(mean_sent = mean(ave_sentiment, na.rm = TRUE),
            std_sent = sd(ave_sentiment, na.rm = TRUE),
            cv_sent = cv(ave_sentiment,na.rm = TRUE))

ggplot( data = concerns_by_month, aes( x=Trig_month, y=value, color=variable)) +
  geom_line(aes(y=mean_sent, color = "mean_sent"), size=2) +
  geom_line(aes(y=std_sent, color = "std_sent"),size =2) +
  ggtitle('Mean Sentiment over months') +
  scale_x_date(labels = date_format("%y-%m"), breaks = "1 month")
```

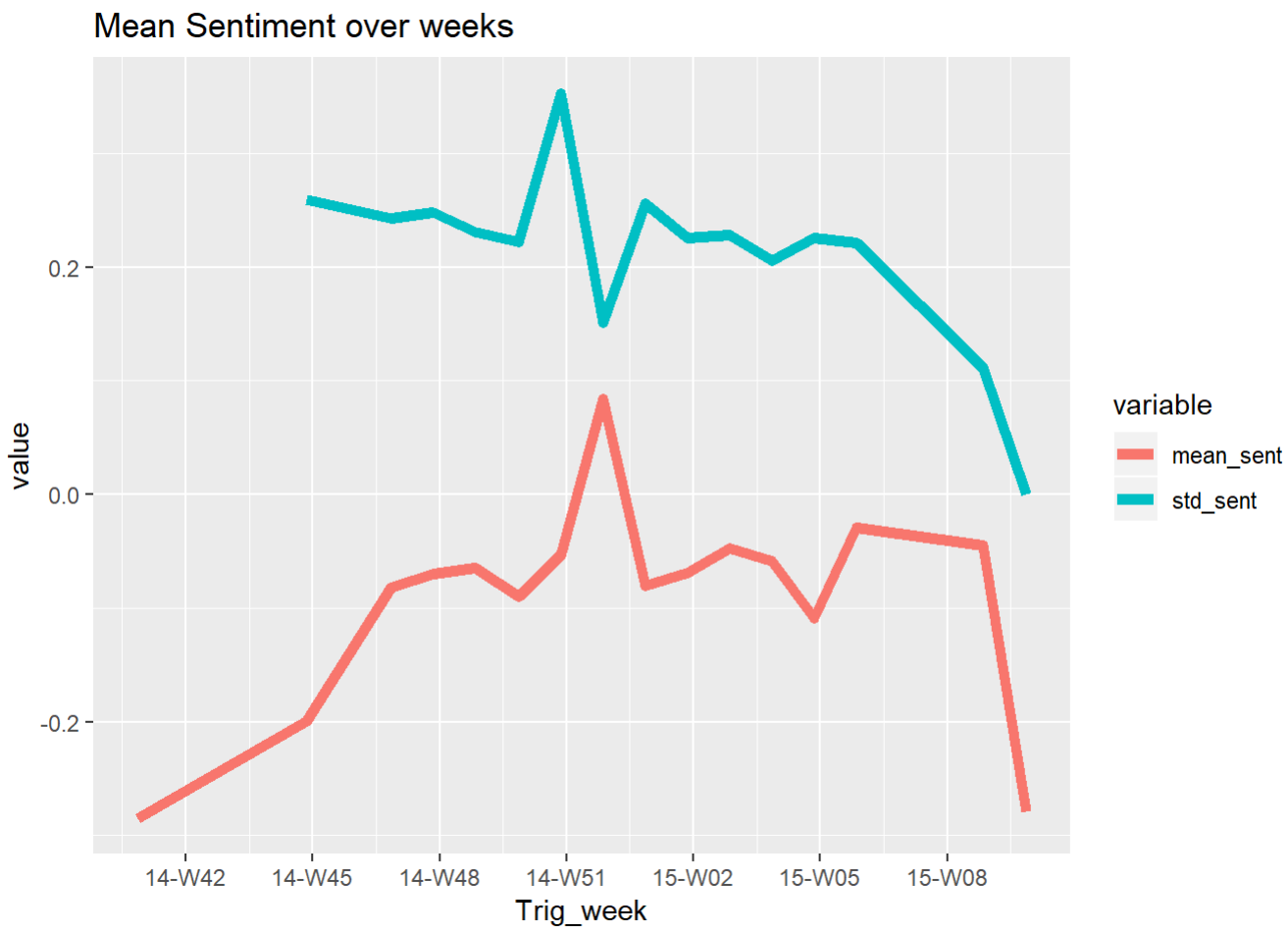
```
## Warning: Removed 1 rows containing missing values (geom_path).
```




## 2.3 Mean Sentiment as a function of week

```
ggplot( data = concerns_by_week, aes( x=Trig_week, y=value, color=variable)) +
  geom_line(aes(y=mean_sent, color = "mean_sent"),size=2) +
  geom_line(aes(y=std_sent, color = "std_sent"),size=2) +
  ggtitle('Mean Sentiment over weeks') +
  scale_x_date(labels = date_format("%y-W%W"), breaks = "3 week")
```

```
## Warning: Removed 1 rows containing missing values (geom_path).
```




# 3 Sentiment by space

```
concerns_by_chiefdom <- sentiment_concerns %>%
  dplyr::select(District, Chiefdom, ave_sentiment)

concerns_by_chiefdom <- concerns_by_chiefdom %>%
  group_by(District,Chiefdom) %>%
  summarise(mean_ave = mean(ave_sentiment, na.rm = TRUE))


concerns_by_district <- sentiment_concerns %>%
  dplyr::select(District, ave_sentiment)

concerns_by_district <- concerns_by_district %>%
  group_by(District) %>%
  summarise(mean_sent = mean(ave_sentiment, na.rm = TRUE),
            sd_sent = sd(ave_sentiment,na.rm = TRUE),
            cv_sent = cv(ave_sentiment,na.rm = TRUE))

counts_by_district <- sentiment_concerns %>%
  dplyr::select(District) %>%
  group_by(District) %>%
  summarise(count_sent = n())

head(concerns_by_district)
```

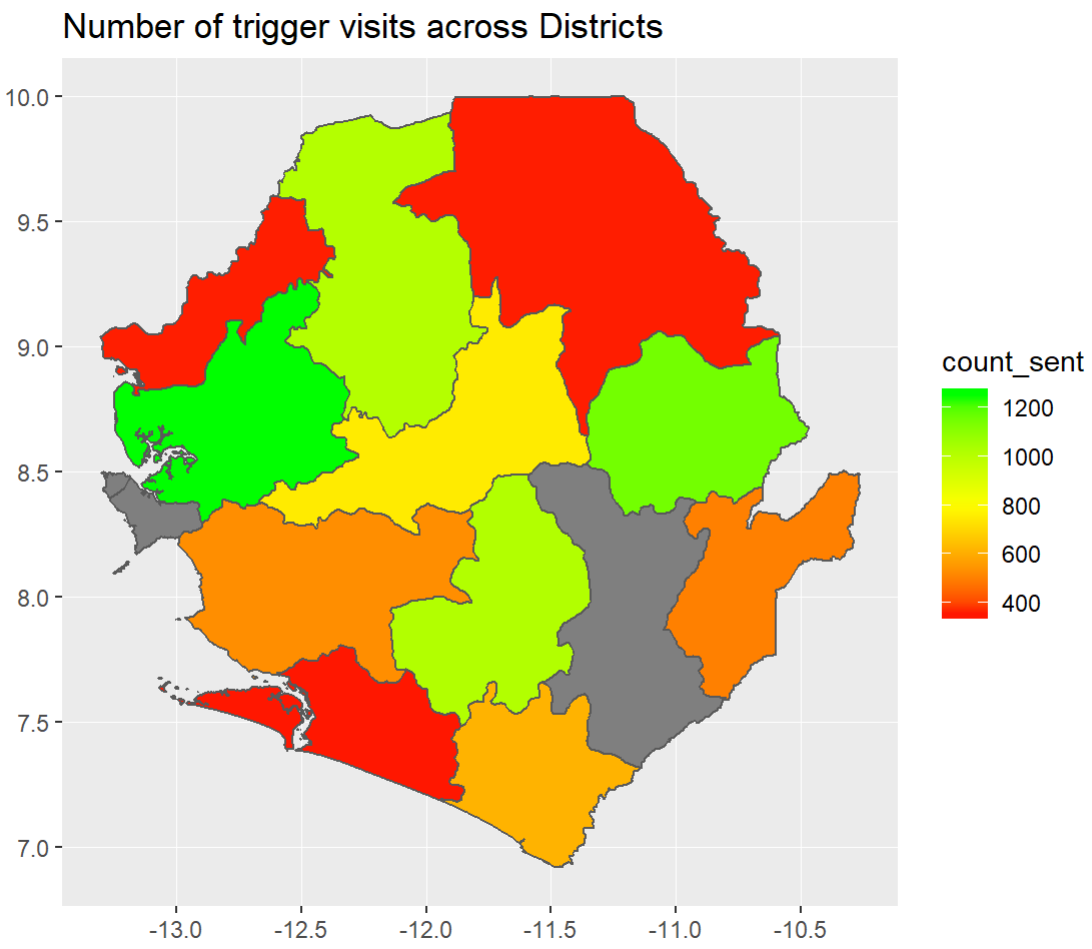
```
## # A tibble: 6 x 4
##   District  mean_sent sd_sent cv_sent
##   <chr>         <dbl>   <dbl>   <dbl>
## 1 Bo          -0.0744   0.218   -293.
## 2 Bombali     -0.0556   0.241   -433.
## 3 Bonthe      -0.109    0.262   -241.
## 4 Kailahun    -0.0451   0.216   -479.
## 5 Kambia      -0.0628   0.222   -353.
## 6 Koinadugu   -0.0547   0.230   -421.
```

# 3.1 Number of trigger visits across districts

```
shp <- read_sf('./shp1/SLE_adm2.shp')
district_df<-data.frame(shp$NAME_2,shp$geometry)

plt <- ggplot()


counts_plot <- district_df %>%
  left_join(counts_by_district,by=c("shp.NAME_2"="District"))
```

```
## Warning: Column `shp.NAME_2`/`District` joining factor and character
## vector, coercing into character vector
```

```
#write.csv(data.frame(shp$admin3Name,shp$admin3RefN,shp$admin2Name),file = "ShapeFile3.csv")

plt + geom_sf(data = counts_plot, aes(fill=count_sent)) + scale_fill_gradient2(low='red',high='g
reen',mid = 'yellow', midpoint = 800) + ggtitle("Number of trigger visits across Districts")
```



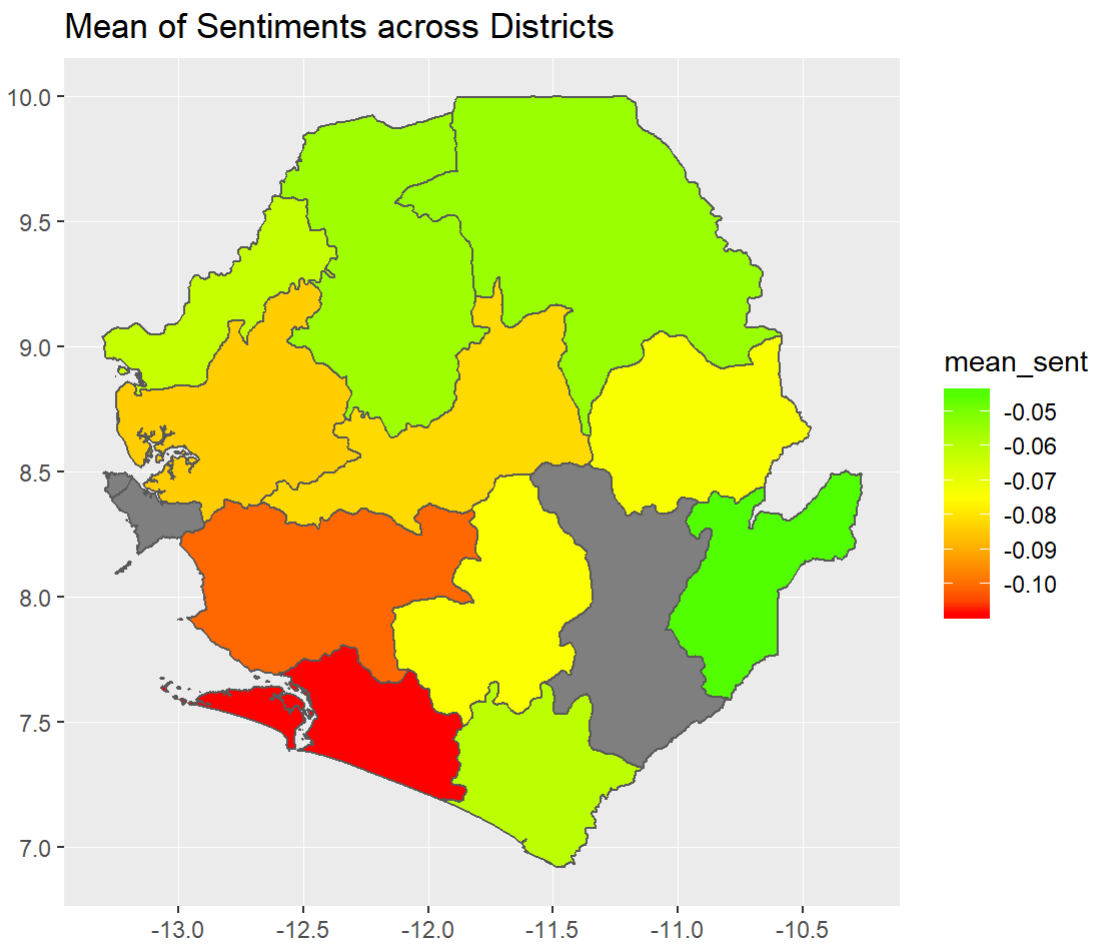

## 3.2 Mean sentiment across districts

Notes and interpretations about sentiment by space: - Freetown and another district missing from the data - south-western part more negative sentiment - Measures against ebola may be better in other parts of the country

```
concerns_plot<-district_df %>%
  left_join(concerns_by_district,by=c("shp.NAME_2"="District"))
```

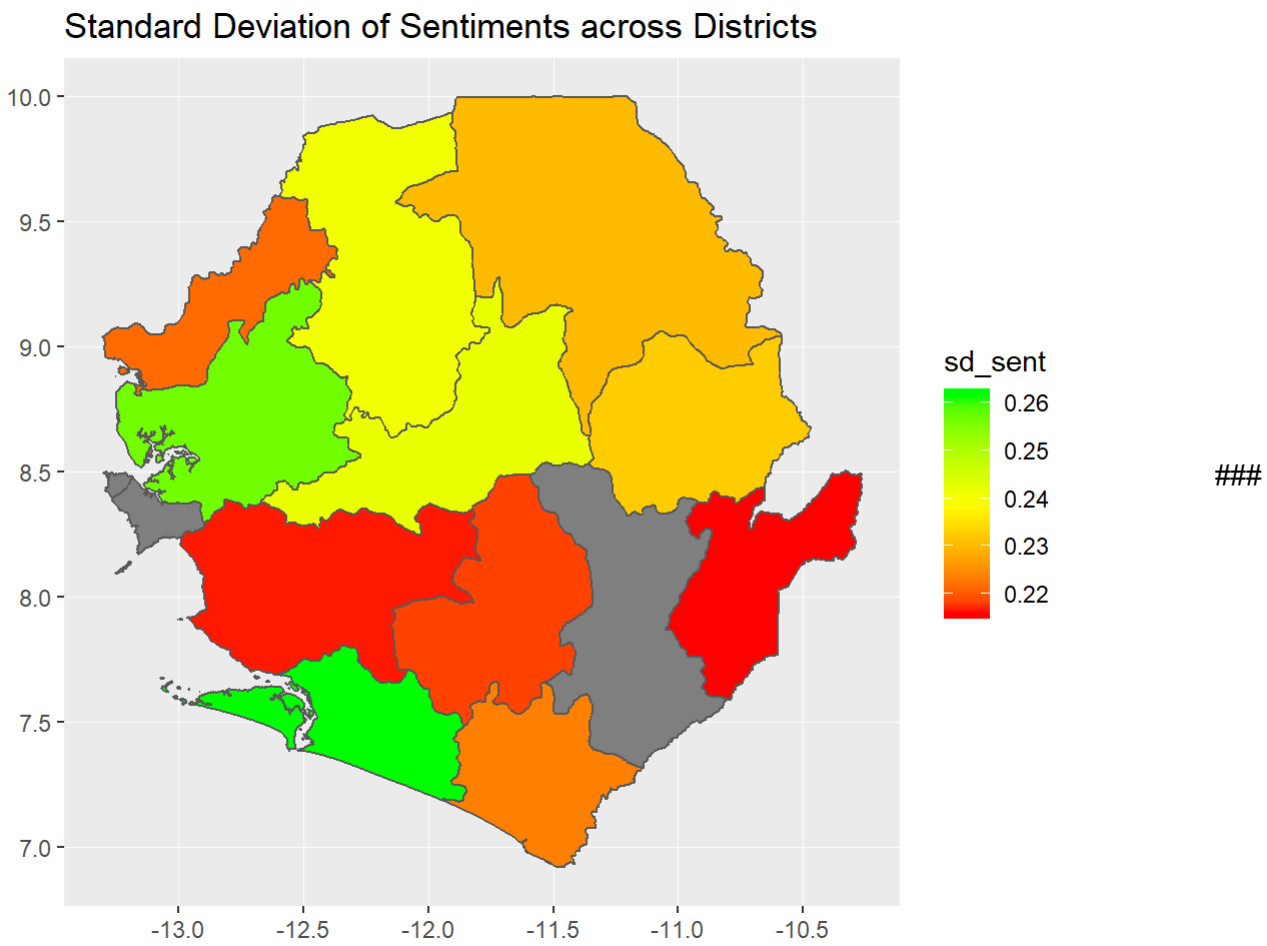
```
## Warning: Column `shp.NAME_2`/`District` joining factor and character
## vector, coercing into character vector
```

```
#write.csv(data.frame(shp$admin3Name,shp$admin3RefN,shp$admin2Name),file = "ShapeFile3.csv")
plt <- ggplot()
plt + geom_sf(data = concerns_plot, aes(fill=mean_sent)) + scale_fill_gradient2(low='red',high=
'green', mid='yellow',midpoint=-0.075) + ggtitle("Mean of Sentiments across Districts")
```




## 3.3 SD of sentiments across districts

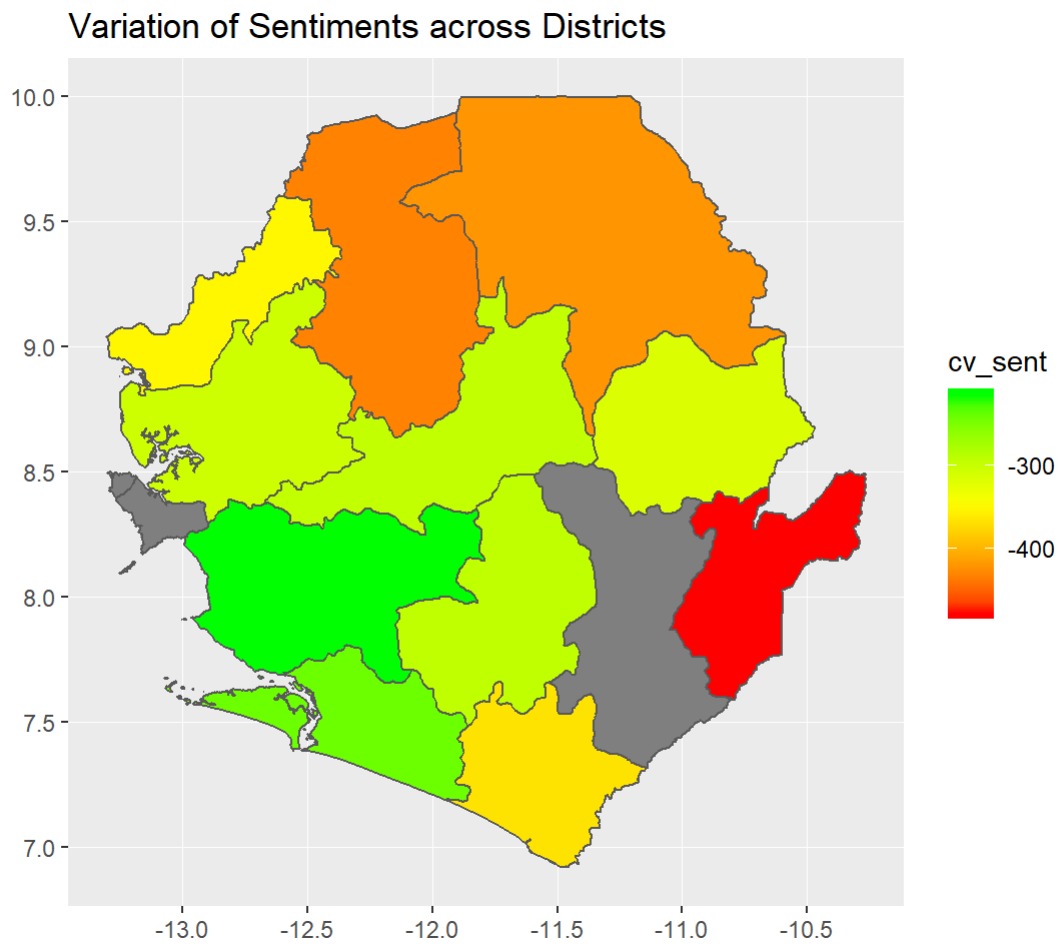
```
sd_max <- max(concerns_plot[!is.na(concerns_plot$sd_sent),]$sd_sent)
sd_min <- min(concerns_plot[!is.na(concerns_plot$sd_sent),]$sd_sent)
sd_mid <- (sd_max+sd_min)/2
plt <- ggplot()
plt + geom_sf(data = concerns_plot, aes(fill=sd_sent)) + scale_fill_gradient2(low='red',high='gr
een', mid='yellow', midpoint=sd_mid) + ggtitle("Standard Deviation of Sentiments across District
s")
```




###

3.4 Co-efficient of variation of sentiments across districts

```
cv_max <- max(concerns_plot[!is.na(concerns_plot$cv_sent),]$cv_sent)
cv_min <- min(concerns_plot[!is.na(concerns_plot$cv_sent),]$cv_sent)
cv_mid <- (cv_max+cv_min)/2
plt <- ggplot()
plt + geom_sf(data = concerns_plot, aes(fill=cv_sent)) + scale_fill_gradient2(low='red',high='gr
een', mid='yellow', midpoint=cv_mid) + ggtitle("Variation of Sentiments across Districts")
```




```
head(concerns_plot)
```

```
##   shp.NAME_2                       geometry   mean_sent   sd_sent
## 1   Kailahun MULTIPOLYGON (((-10.30196 8... -0.04505028 0.2159075
## 2     Kenema MULTIPOLYGON (((-11.49417 8...          NA        NA
## 3       Kono MULTIPOLYGON (((-11.03017 9... -0.07359921 0.2323381
## 4    Bombali MULTIPOLYGON (((-11.90307 9... -0.05558179 0.2406574
## 5     Kambia MULTIPOLYGON (((-13.13486 8... -0.06281015 0.2215894
## 6  Koinadugu MULTIPOLYGON (((-11.20397 9... -0.05468942 0.2303602
##     cv_sent
## 1 -479.2589
## 2        NA
## 3 -315.6802
## 4 -432.9789
## 5 -352.7924
## 6 -421.2152
```

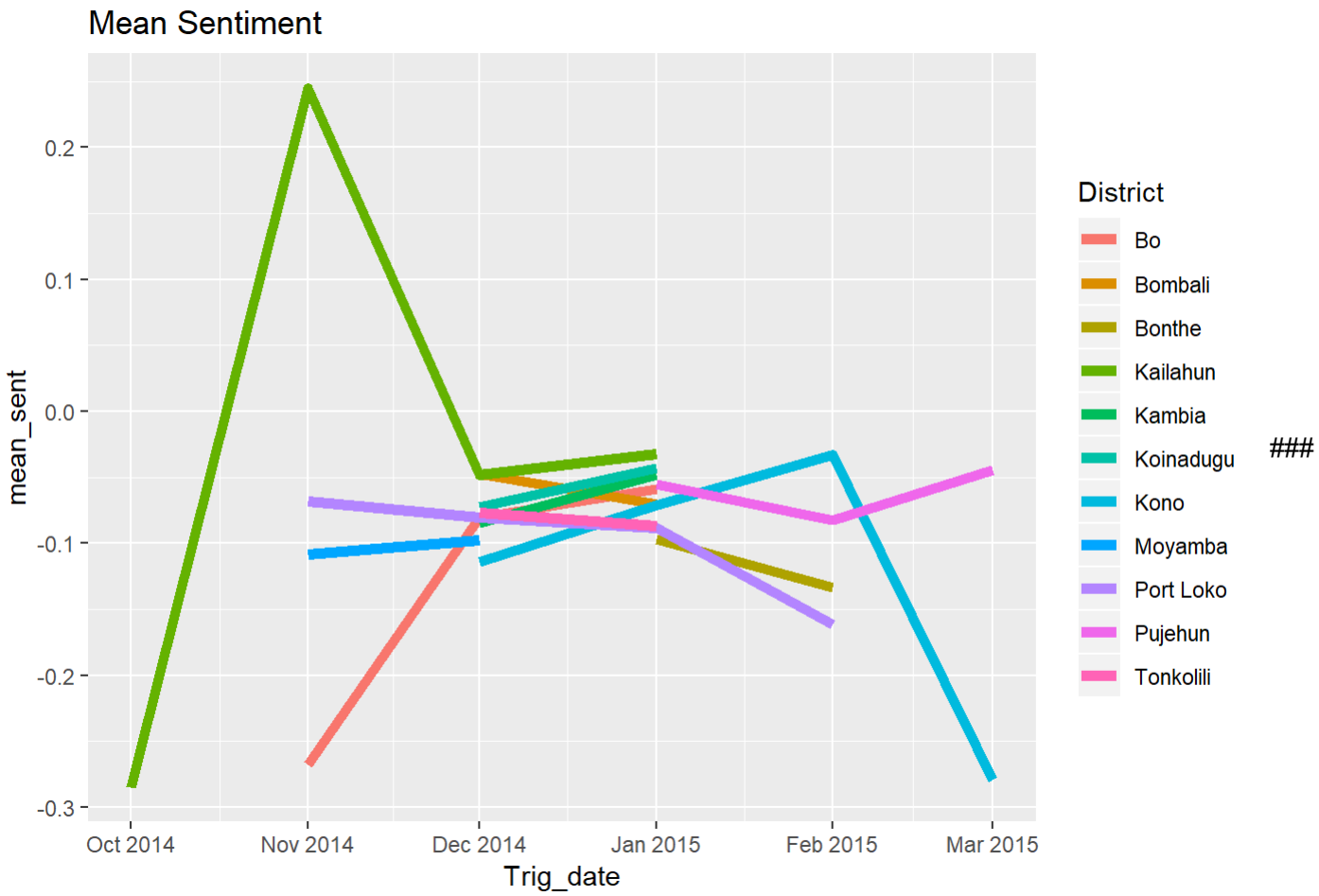
# 4 Sentiment by time and space

## 4.1 Mean sentiment by months across districts

```
concerns_dis_mon <- sentiment_concerns %>%
  dplyr::select(Trig_date,District, ave_sentiment)

concerns_dis_mon$Trig_date <- as.Date(as.yearmon(concerns_dis_mon$Trig_date))
#concerns_dis_mon$Trig_date <- as.Date(concerns_dis_mon$Trig_date, "%b %Y")

concerns_dis_mon <- concerns_dis_mon %>%
  group_by(District,Trig_date) %>%
  summarise(mean_sent = mean(ave_sentiment, na.rm = TRUE))

concerns_dis_mon
```

```
## # A tibble: 30 x 3
## # Groups:   District [11]
##    District Trig_date  mean_sent
##    <chr>    <date>         <dbl>
##  1 Bo       2014-11-01   -0.268
##  2 Bo       2014-12-01   -0.0802
##  3 Bo       2015-01-01   -0.0589
##  4 Bombali  2014-12-01   -0.0473
##  5 Bombali  2015-01-01   -0.0705
##  6 Bonthe   2015-01-01   -0.0969
##  7 Bonthe   2015-02-01   -0.133
##  8 Kailahun 2014-10-01   -0.285
##  9 Kailahun 2014-11-01    0.245
## 10 Kailahun 2014-12-01   -0.0483
## # ... with 20 more rows
```

```
ggplot( data = concerns_dis_mon, aes(x=Trig_date, y=mean_sent)) + geom_line(aes(color = Distric
t), size=2) + ggtitle('Mean Sentiment') + scale_x_date(name="Trig_date", date_labels = "%b %Y",
 breaks="month" )
```

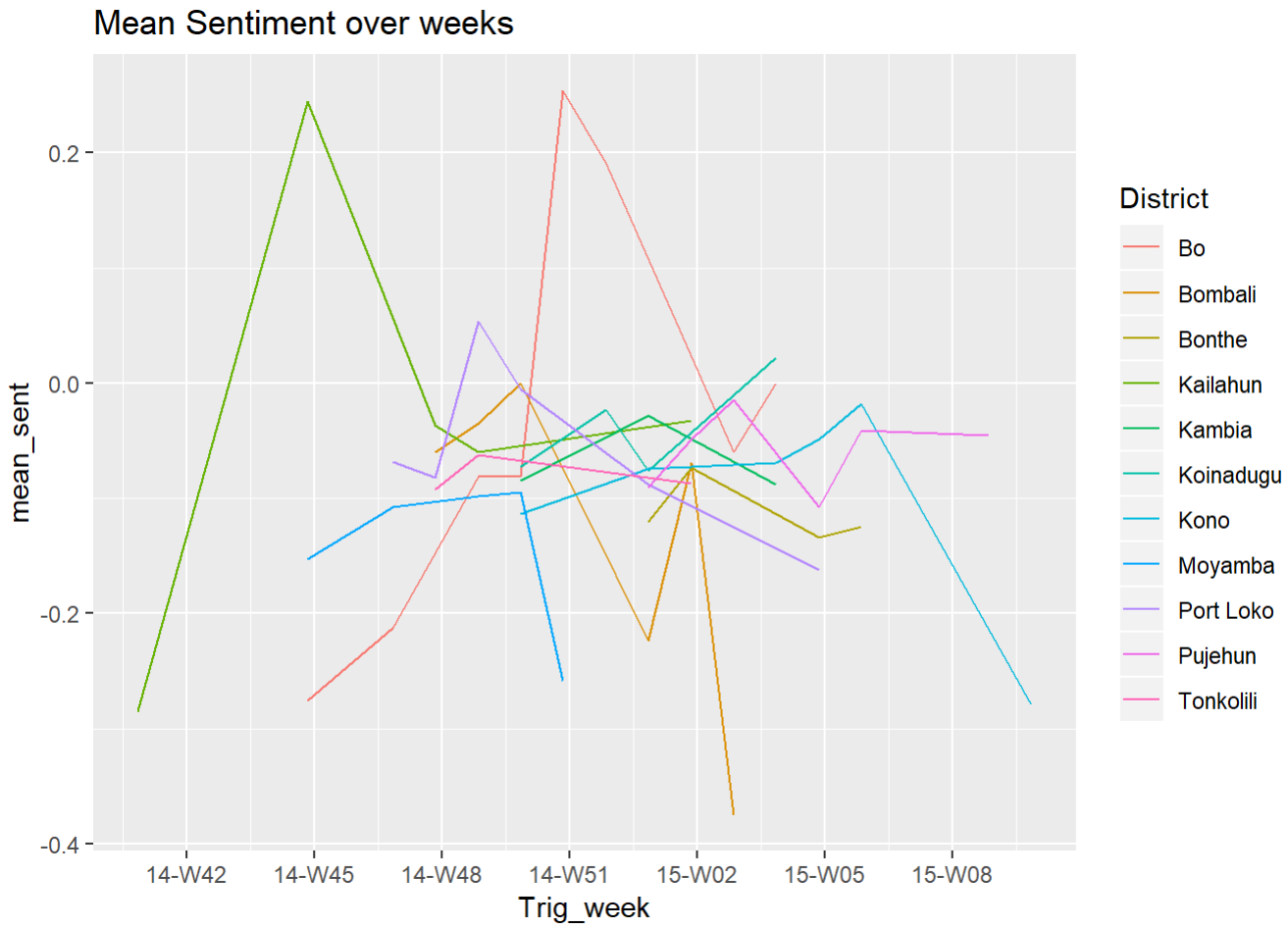



4.2 Mean sentiment by weeks across districts

```
concerns_dis_week <- sentiment_concerns %>%
  dplyr::select(Trig_date,District, ave_sentiment)

concerns_dis_week$Trig_week <- as.Date(cut(concerns_dis_week$Trig_date,breaks = "week",start.on.
monday = FALSE))

concerns_dis_week <- concerns_dis_week %>%
  group_by(District,Trig_week) %>%
  summarise(mean_sent = mean(ave_sentiment, na.rm = TRUE))

concerns_dis_week
```

```
## # A tibble: 56 x 3
## # Groups:   District [11]
##    District Trig_week  mean_sent
##    <chr>    <date>         <dbl>
##  1 Bo       2014-11-09   -0.276
##  2 Bo       2014-11-23   -0.212
##  3 Bo       2014-12-07   -0.0813
##  4 Bo       2014-12-14   -0.0810
##  5 Bo       2014-12-21    0.254
##  6 Bo       2014-12-28    0.191
##  7 Bo       2015-01-18   -0.0598
##  8 Bo       2015-01-25    0
##  9 Bombali  2014-11-30   -0.0594
## 10 Bombali  2014-12-07   -0.0349
## # ... with 46 more rows
```

```
ggplot( data = concerns_dis_week, aes(x=Trig_week, y=mean_sent)) + geom_line(aes(color = District)) + ggtitle('Mean Sentiment over weeks') + scale_x_date(name="Trig_week", date_labels = "%y-W%W", breaks="3 week" )
```




# 4.3 Mean sentiment by months across district (Animation)

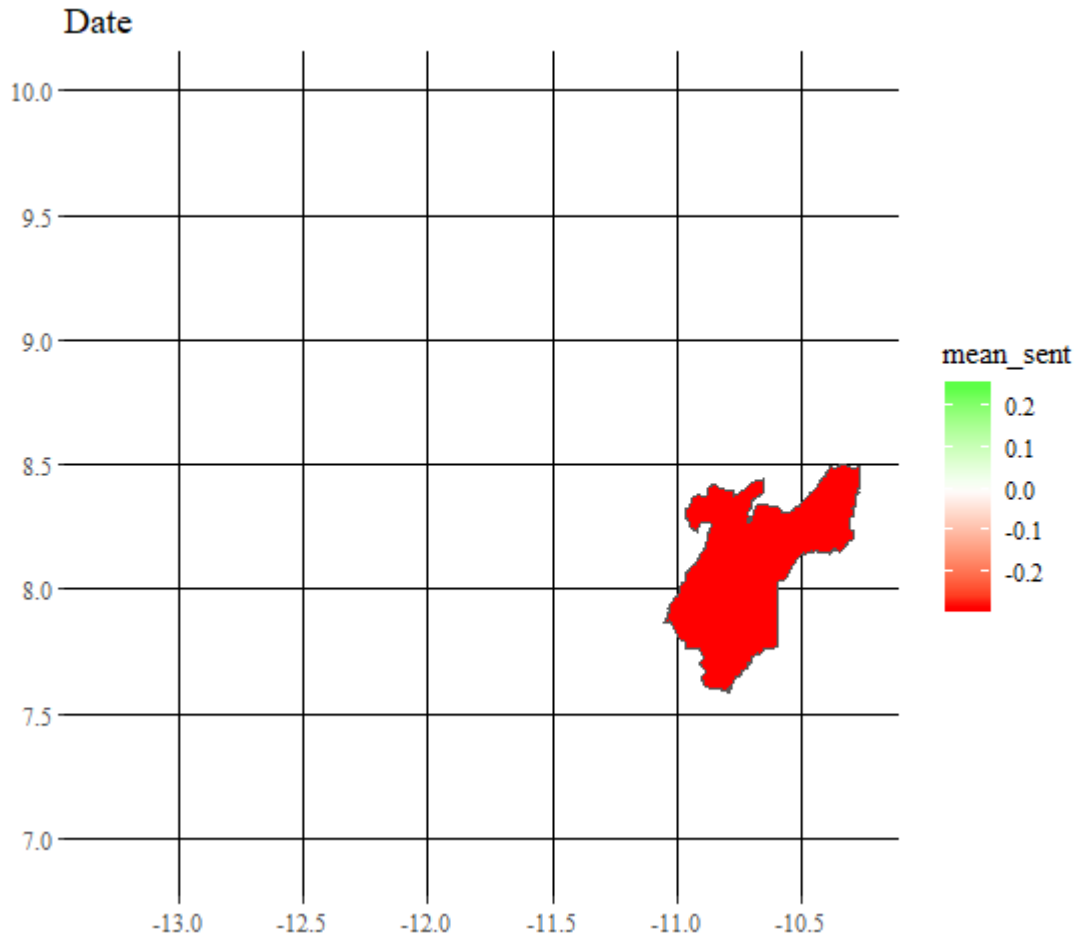
```
concerns_month_geo <-district_df %>%
  left_join(concerns_dis_mon,by=c("shp.NAME_2"="District"))
```

```
## Warning: Column `shp.NAME_2`/`District` joining factor and character
## vector, coercing into character vector
```

```
concerns_month_geo <- concerns_month_geo[!is.na(concerns_month_geo$Trig_date),]


concerns_month_geo_animate <- ggplot() + geom_sf(data = concerns_month_geo, aes(fill=mean_sent,f
rame=Trig_date)) + scale_fill_gradient2(low='red',high='green', mid='white',midpoint=0) + transi
tion_states(Trig_date, wrap=TRUE) + coord_sf() + theme_tufte() + labs(title = "Date")
```

```
## Warning: Ignoring unknown aesthetics: frame
```

```
## animate plot with gganimate
animate(concerns_month_geo_animate, fps=3)
```




```
#anim_save("animation.gif",animation=last_animation())

#concerns_month_geo$Trig_month

#concerns_month_geo
```