

# Twitter EDA

Kristin Köhler

25/09/2023

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
```

```
library(tidyverse)
library(wordcloud2)
library(tidytext)
library(data.table)
```

```
Sys.setlocale(category="LC_ALL", locale="German")
```

```
## Warning in Sys.setlocale(category = "LC_ALL", locale = "German"): OS reports
## request to set locale to "German" cannot be honored
```

Read and process all Twitter files (tweets and user data)

```
library(rjson)

## tweet data
tweets = list()
for (file in list.files("/Users/kristinkoehler/Desktop/NLP_DL/long_covid2/twitter/",
  pattern = "data*", recursive = TRUE)) {
  print(file)

  tweets[[file]] = jsonlite::fromJSON(txt = paste0("/Users/kristinkoehler/Desktop/NLP_DL/long_covid2/twitter/",
    file), flatten = TRUE)
  tweets[[file]] = tweets[[file]] %>%
```

```

      select(text, id, created_at, author_id, public_metrics.retweet_count,
             public_metrics.reply_count, public_metrics.like_count,
             public_metrics.quote_count, public_metrics.impression_count,
             entities.mentions, entities.hashtags, entities.urls)
    }
  tweets_merged = bind_rows(tweets)
  tweets_merged = tweets_merged %>%
    unnest_wider(c(entities.mentions, entities.hashtags), names_sep = ".")
  saveRDS(tweets_merged, "TweetDataMerged.rds")

## user data
users_info = list()
for (file in list.files("/Users/kristinkoehler/Desktop/NLP_DL/long_covid2/twitter/",
  pattern = "user*", recursive = TRUE)) {
  print(file)

  users_info[[file]] = jsonlite::fromJSON(txt = paste0("/Users/kristinkoehler/Desktop/NLP_DL/long_covid2/twitter/",
    file), flatten = TRUE)$users
}
users_info = reduce(users_info, dplyr::full_join)
saveRDS(users_info, "UserDataMerged.rds")

```

```

tweets_merged <- readRDS("/Users/kristinkoehler/Desktop/NLP_DL/TweetDataMerged.rds")
users_info <- readRDS("/Users/kristinkoehler/Desktop/NLP_DL/UserDataMerged.rds")

```

Tweet information

```
names(tweets_merged)
```

```

## [1] "text"
## [3] "created_at"
## [5] "public_metrics.retweet_count"
## [7] "public_metrics.like_count"
## [9] "public_metrics.impression_count"
## [11] "entities.mentions.end"

      "id"
      "author_id"
      "public_metrics.reply_count"
      "public_metrics.quote_count"
      "entities.mentions.start"
      "entities.mentions.username"

```

```
## [13] "entities.mentions.id"          "entities.hashtags.start"
## [15] "entities.hashtags.end"        "entities.hashtags.tag"
## [17] "entities.urls"
```

```
cat("#tweets: ", nrow(tweets_merged))
```

```
## #tweets: 1006594
```

```
cat("#duplicated tweets: ", sum(duplicated(tweets_merged$text)))
```

```
## #duplicated tweets: 757543
```

```
as.data.frame(table(tweets_merged$text)) %>%
  arrange(-Freq) %>%
  head(n = 10)
```

```
##
```

```
## 1 RT @IchBin_R0: Kennt ihr persönlich jemanden, der an #LongCovid leidet? \n\nRT = ja\n\n
## 2 RT @Karl_Lauterbach: Noch nie war die Inzidenz von Kindern und Jugendlichen so hoch wie heute. Da Eltern und zT auch Lehrpersonal no
## 3 RT @Karl_Lauterbach: Im nächsten Jahr werden hunderttausende in Deutschland mit den Folgen von #LongCovid kämpfen. Sie werden nicht
## 4 RT @Karl_Lauterbach: Jetzt führt @c_lindner bei #AnneWill die „enormen wirtschaftlichen Folgen“ der Notbremse an. Was kosten Ausgang
## 5 RT @TimoStrotmann: Das bin ich mit meine Frau. Vor einem Jahr waren wir noch glücklich, gesund und hatten großen Plänen für unsere Z
## 6 RT @Karl_Lauterbach: Spiegel Autorin @marga_owski ist leider #LongCovid erkrankt. Besser kann man leider nicht beschreiben, wie si
## 7 RT @Karl_Lauterbach: Hier die Übersicht der Inzidenz 5-14 J für ganz Deutschland. Die Lage in NRW ist ohne Beispiel. Kinder sollten
## 8 RT @Karl_Lauterbach: Das haben die Kinder nicht verdient. Sie haben die Erwachsenen geschützt, Opfer gebracht, und sollen jetzt ohne
## 9 RT @quervain_de: COVID-19 führt auch nach mildem Verlauf häufig (10% plus) zu anhaltenden neurologischen Störungen wie Gedächtnispro
## 10 RT @Karl_Lauterbach: Hier mein Thread zum Asthmaspray Budesonid. 3. Tage nach Symptombeginn genommen sinkt die Wahrscheinlichkeit ei
```

```
## Freq
## 1 2533
## 2 2169
## 3 1540
## 4 1533
## 5 1485
## 6 1484
## 7 1405
## 8 1342
## 9 1342
## 10 1327
```

User information

```
nrow(users_info)
```

```
## [1] 209628
```

```
names(users_info)
```

```
## [1] "name"                "protected"
## [3] "verified"            "username"
## [5] "description"         "profile_image_url"
## [7] "created_at"          "id"
## [9] "location"            "pinned_tweet_id"
## [11] "url"                 "public_metrics.followers_count"
## [13] "public_metrics.following_count" "public_metrics.tweet_count"
## [15] "public_metrics.listed_count" "entities.description.hashtags"
## [17] "entities.description.urls" "entities.description.mentions"
## [19] "entities.url.urls" "entities.description.cashtags"
## [21] "withheld.country_codes" "withheld.scope"
```

```
length(unique(users_info$id))
```

```
## [1] 98916
```

The author IDs fully intersect with the user metadata

```
all(tweets_merged$author_id %in% users_info$id)
```

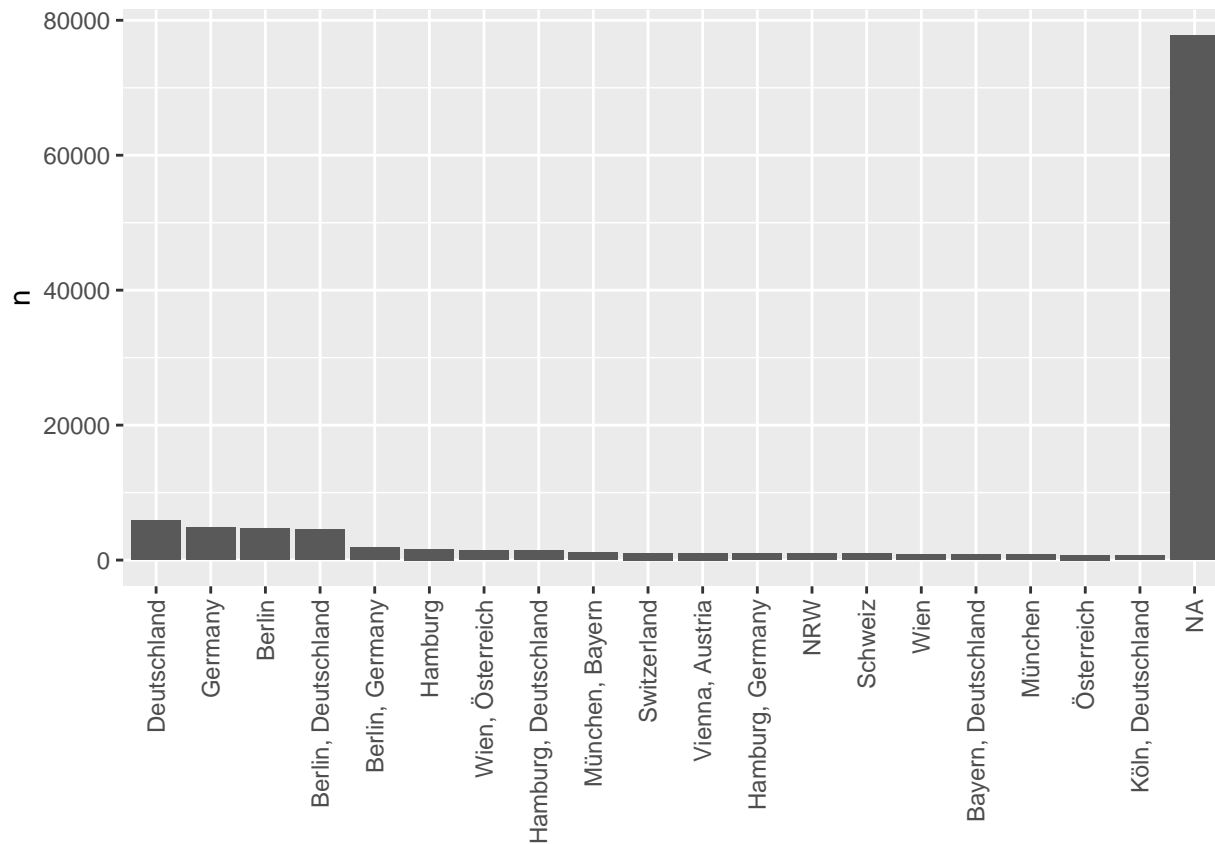
```
## [1] TRUE
```

User locations

```
top_locations <- users_info %>%
  count(location) %>%
  arrange(-n) %>%
  top_n(n = 20)
```

```
## Selecting by n
```

```
top_locations$location <- factor(top_locations$location, levels = top_locations$location)
ggplot(top_locations) + geom_bar(aes(x = location, y = n), stat = "identity",
  position = "dodge") + theme(axis.text.x = element_text(angle = 90,
  vjust = 0.5, hjust = 1), axis.title.x = element_blank())
```



Some other statistics

```
summary(users_info %>%
  select_if(is.numeric))
```

```
## public_metrics.followers_count public_metrics.following_count
## Min.      :      -1          Min.      :      -1
## 1st Qu.:    113            1st Qu.:   235
## Median :    464            Median :   582
## Mean   :  154558          Mean    :  1204
## 3rd Qu.:   1799            3rd Qu.:  1339
## Max.    :127042690         Max.     :976703
##
## public_metrics.tweet_count public_metrics.listed_count
## Min.      :      -1          Min.      :    0.0
## 1st Qu.:   2701            1st Qu.:    0.0
## Median :  10682            Median :    4.0
## Mean   :   36262          Mean     :   446.1
## 3rd Qu.:  34620            3rd Qu.:   25.0
## Max.    :3343413          Max.     :216581.0
##
##                               NA's      :1
```

Number of tweets per user

```
tweets_merged %>%
  plyr::count("author_id") %>%
  merge(users_info[, c("username", "id")], by.x = "author_id",
    by.y = "id") %>%
  distinct() %>%
  arrange(-freq) %>%
  top_n(n = 20, wt = freq)
```

```
##          author_id freq      username
## 1          391576085 7325      wunderwer
## 2 1283081430017482756 6123      OStriebling
## 3 1374012430523715587 5236 Arizona10238417
## 4 1457007704111820804 4140 charlotta123456
## 5  832377641059680257 3765      Hervorrager
## 6 1272451027892125696 3506      KippelsNicole
```

```
## 7 1354697491392966656 3464 LangzeitC
## 8 1238191614092292097 3355 BettinafightsME
## 9 112526774 2793 Michael_Kunz
## 10 1456759321166032898 2753 radineulla
## 11 1225431487253753857 2549 GreenSaxo
## 12 1115385542 2501 symptomodyssee
## 13 831448883628736512 2476 Jo_Hanna_Jo
## 14 857972066980687872 2454 Hopps25
## 15 2403082631 2451 suse1603
## 16 1365696248284647424 2344 printi71
## 17 404377878 2306 IchBin_R0
## 18 869501900 2164 AutOdated
## 19 1378008150926884868 2119 uLAUJzaboci4NSj
## 20 4255832955 2055 goddyw567
```

Top 20 most retweeted tweets

```
mostRetweeted_tweets_top20 <- tweets_merged[!startsWith(tweets_merged$text,
  "RT"), ]
mostRetweeted_tweets_top20 <- (mostRetweeted_tweets_top20 %>%
  arrange(-public_metrics.retweet_count))[1:20, c("text", "author_id",
  "public_metrics.retweet_count")]
mostRetweeted_tweets_top20 <- merge(mostRetweeted_tweets_top20,
  unique(users_info[c("id", "username")] ), by.x = "author_id",
  by.y = "id")
mostRetweeted_tweets_top20[c("text", "username")]
```

##

## 1

COVID-19 führt auch nach mildem Verlauf häufig (10% plus)

## 2

## 3 Jetzt warnt auch die @WHO explizit vor aktivierender #Reha bei #LongCovid , wenn #PEM bzw #PESE vorliegt!\n\nVielleicht

## 4 Die Genußtuung, ja fast Schadenfreude über die Schwierigkeit, #Delta in #Australien durch #NoCovid einzudämmen,

## 5 Noch nie war die Inzidenz von Kindern und Jugendlichen so hoch wie heute. Da Eltern und zT auch Lehrpersonal noch nicht geimpft

## 6 Jetzt führt @c\_lindner bei #AnneWill die „enormen wirtschaftlichen Folgen" der Notbremse an. Was kosten

## 7 Im nächsten Jahr werden hunderttausende in Deutschland mit den Folgen von

## 8 Spiegel Autorin @marga\_owski ist leider #LongCovid erkrankt. Besser kann man leider nicht beschreiben

## 9 Hier mein Thread zum Asthmaspray E

## 10 Hier die Über

14 J für ganz Deutschland. Die Lage in NRW ist ohne Beispiel. Kinder sollten nicht dem Risiko von #LongCovid ausgesetzt werden. <https://t.>

```

## 11                Mindestens 10% Covid Infizierten entwickeln #LongCovid. Die meisten sind 6 Monate später nicht symptomfrei
## 12                                                    Das haben die Kinder
## 13 Es ist höchste Zeit dass #LongCovid für Kinder Thema in Deutschland wird. Wir stehen vielleicht vor Durchseuchung einer ganzen Generation
## 14                (1) Artikel ist traurig aber nicht falsch. Wenn Ältere und Risikopatienten in nächsten Monaten geimpft werden
## 15                Was für ein Kitsch. Von der FDP kommt einfach immer nur Druck, die Geschäfte zu öffnen. Man sagt „Freiheit“ und meint „Umsatz“
## 16
## 17                Das bin ich mit meine Frau. Vor einem Jahr waren wir noch glücklich, gesund und hatten Kinder
## 18 #COVID19 - „Unheimliche Ähnlichkeit mit HIV“\n\nDr. Grace McComsey, eine Spezialistin für Infektionskrankheiten an den Universitätsklinikum
## 19                Ich habe mal etwas Zeit investiert und die folgende aktuelle (13.1.) Meta-Studie zu #LongCovid durchgearbeitet
Link am Ende.\n\n1/n
## 20                Eine geschätzte Kollegin unserer ICU ist endgültig arbeitsunfähig, sie wird nie mehr wiederkommen. Sie hat Kinder
##                username
## 1                quervain_de
## 2                RenzPolster
## 3                katha1970
## 4                EckerleIsabella
## 5                Karl_Lauterbach
## 6                Karl_Lauterbach
## 7                Karl_Lauterbach
## 8                Karl_Lauterbach
## 9                Karl_Lauterbach
## 10               Karl_Lauterbach
## 11               Karl_Lauterbach
## 12               Karl_Lauterbach
## 13               Karl_Lauterbach
## 14               Karl_Lauterbach
## 15               Karl_Lauterbach
## 16                IchBin_R0
## 17                TimoStrotmann
## 18                RWittenbrink
## 19                olewin
## 20                narkosedoc

```

Number of tweets per year/month

```

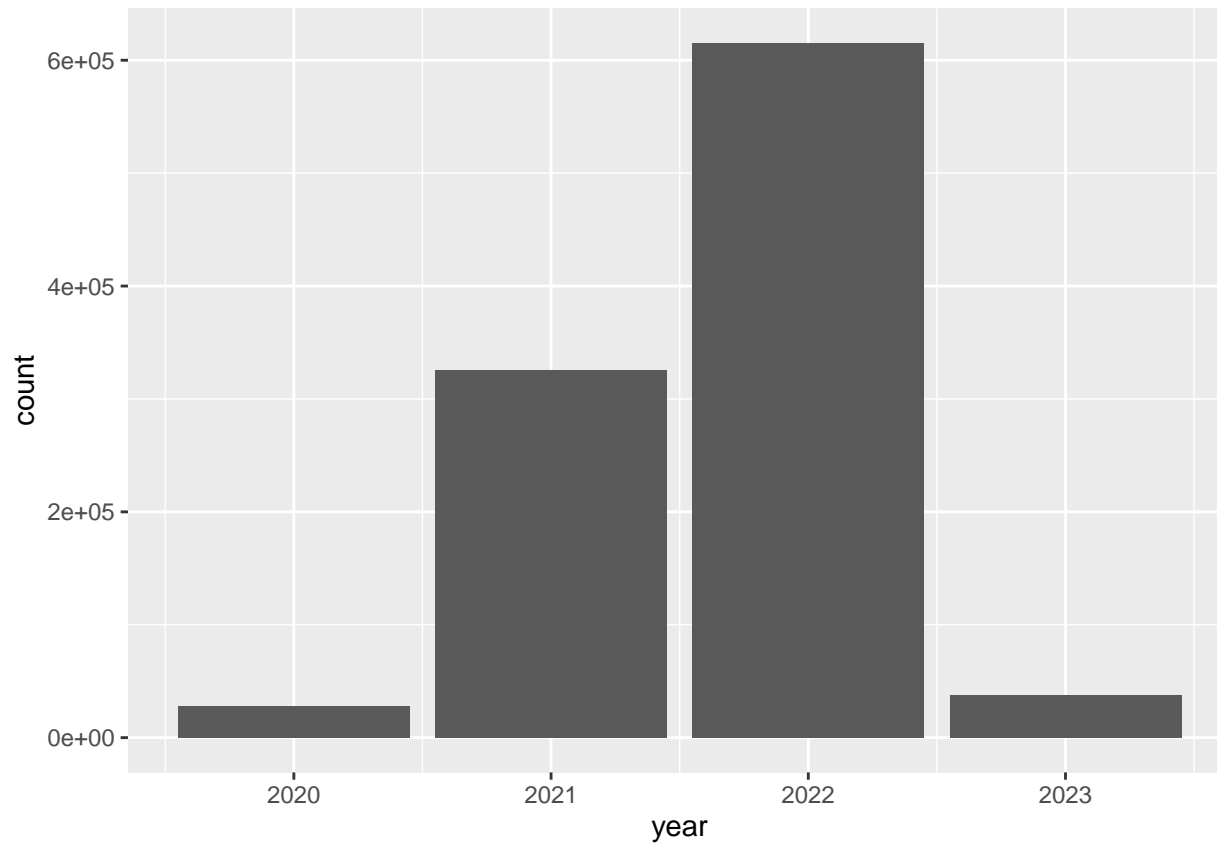
tweets_merged$created_at = as.Date(tweets_merged$created_at)
tweets_merged$year = year(tweets_merged$created_at)

texts = tweets_merged[c("text", "created_at", "year")]

```



```
texts %>%  
  ggplot() + geom_bar(aes(x = year))
```



```
texts <- data.table(texts)  
texts_counts <- texts[, .N, by = .(year(created_at), month(created_at),  
  day(created_at))]  
texts_counts$date <- paste0(texts_counts$year, "-", texts_counts$month,  
  "-", texts_counts$day)  
texts_counts$date <- as.Date(texts_counts$date, format = "%Y-%m-%d")
```

Plot number of Covid infections ([https://www.rki.de/DE/Content/InfAZ/N/Neuartiges\\_Coronavirus/Daten/Inzidenz-Tabellen.html](https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Daten/Inzidenz-Tabellen.html))

```
library("readxl")
fallzahlen23 <- read_excel("/Users/kristinkoehler/Desktop/NLP_DL/Fallzahlen_2021-23.xlsx",
  sheet = 3)
fallzahlen20 <- read_excel("/Users/kristinkoehler/Desktop/NLP_DL/Fallzahlen_2020-21.xlsx",
  sheet = 2)

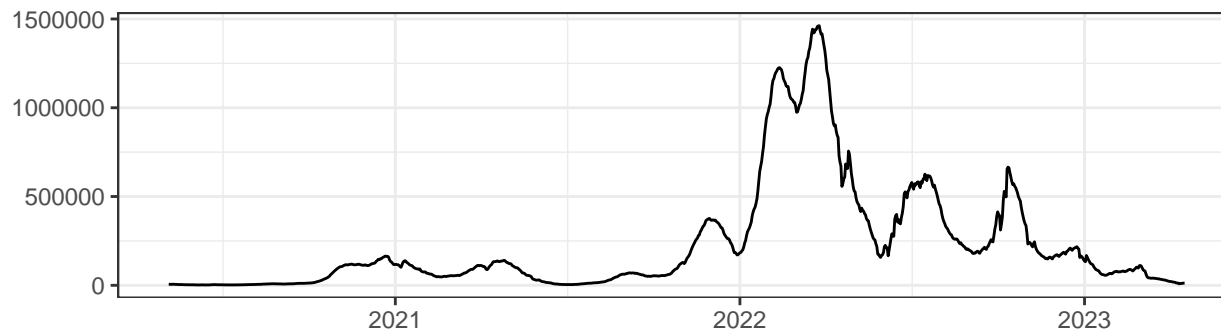
fallzahlen_gesamt <- data.frame(dates = c(colnames(fallzahlen20)[-1],
  colnames(fallzahlen23)[-1]), infectionNr = c(as.numeric(fallzahlen20[nrow(fallzahlen20),
  -1]), as.numeric(fallzahlen23[nrow(fallzahlen23), -1])))
fallzahlen_gesamt$dates <- str_replace_all(fallzahlen_gesamt$dates,
  "2021", "21")
fallzahlen_gesamt$dates <- str_replace_all(fallzahlen_gesamt$dates,
  "2022", "22")
fallzahlen_gesamt$dates <- str_replace_all(fallzahlen_gesamt$dates,
  "2023", "23")
fallzahlen_gesamt$dates <- as.Date(fallzahlen_gesamt$dates, format = "%d.%m.%y")

fallzahlen_p <- ggplot(fallzahlen_gesamt) + geom_line(aes(x = dates,
  y = infectionNr)) + xlab("") + ggtitle("Number of Covid Infections") +
  theme_bw() + ylab("")

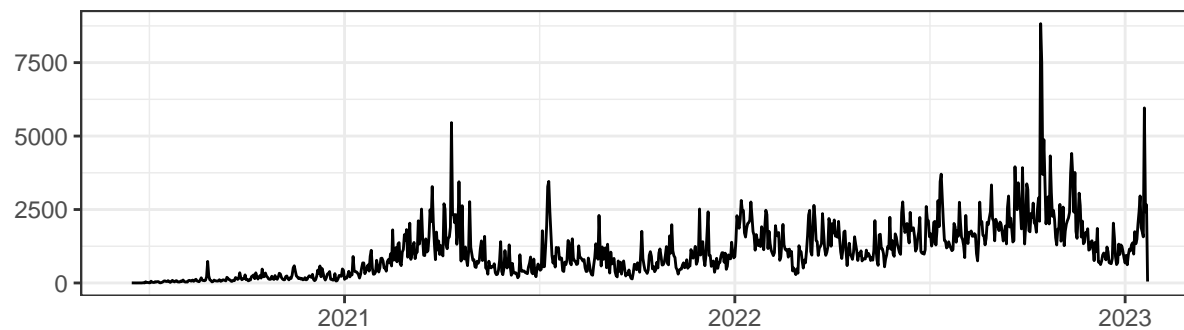
timeline_tweets <- texts_counts %>%
  ggplot() + geom_line(aes(x = date, y = N)) + ggtitle("Number of Tweets related to Long-Covid") +
  xlab("") + ylab("") + theme_bw()

library(grid)
grid.newpage()
grid.draw(rbind(ggplotGrob(fallzahlen_p), ggplotGrob(timeline_tweets),
  size = "last"))
```

Number of Covid Infections



Number of Tweets related to Long-Covid



Read stop words file

```
de_stop_words <- read_lines("/Users/kristinkoehler/Desktop/NLP_DL/stopwords-de.txt")
```

Most commonly used words in tweets (without stop words)

```
##count and plot word frequencies
lead_words <- texts %>%
  unnest_tokens(word, text, token = "words") %>%
  filter(!word %in% de_stop_words) %>% # delete stop words
  filter(nchar(word) >= 3) %>% # 3 letters
```

```

filter(word!="https")%>%
filter(word!="t.co") %>%
filter(!str_detect(word, "[:digit:]" ))

```

```
lead_words_stem <- lead_words
```

```

lead_words_cloud <- lead_words_stem %>%
  dplyr::count(word) %>%
  ungroup() %>%
  dplyr::rename(freq = n) %>%
  arrange(-freq) %>%
  top_n(100)

```

## Selecting by freq

```

top_1000_words = lead_words_stem %>%
  dplyr::count(word) %>%
  ungroup() %>%
  mutate(word = factor(word, levels=word)) %>%
  filter(!is.na(word)) %>%
  arrange(-n) %>%
  top_n(1000, n)

```

```
#write.table(top_1000_words, 'top1000_words.csv', sep = ';')
```

Word cloud

```
# library(htmlwidgets) library(webshot)
```

```

wc <- wordcloud2(lead_words_cloud) #,
# max.words = 100, scale = c(2.4, 0.6), color =
# RColorBrewer::brewer.pal(11, 'Spectral')[1:11])
wc

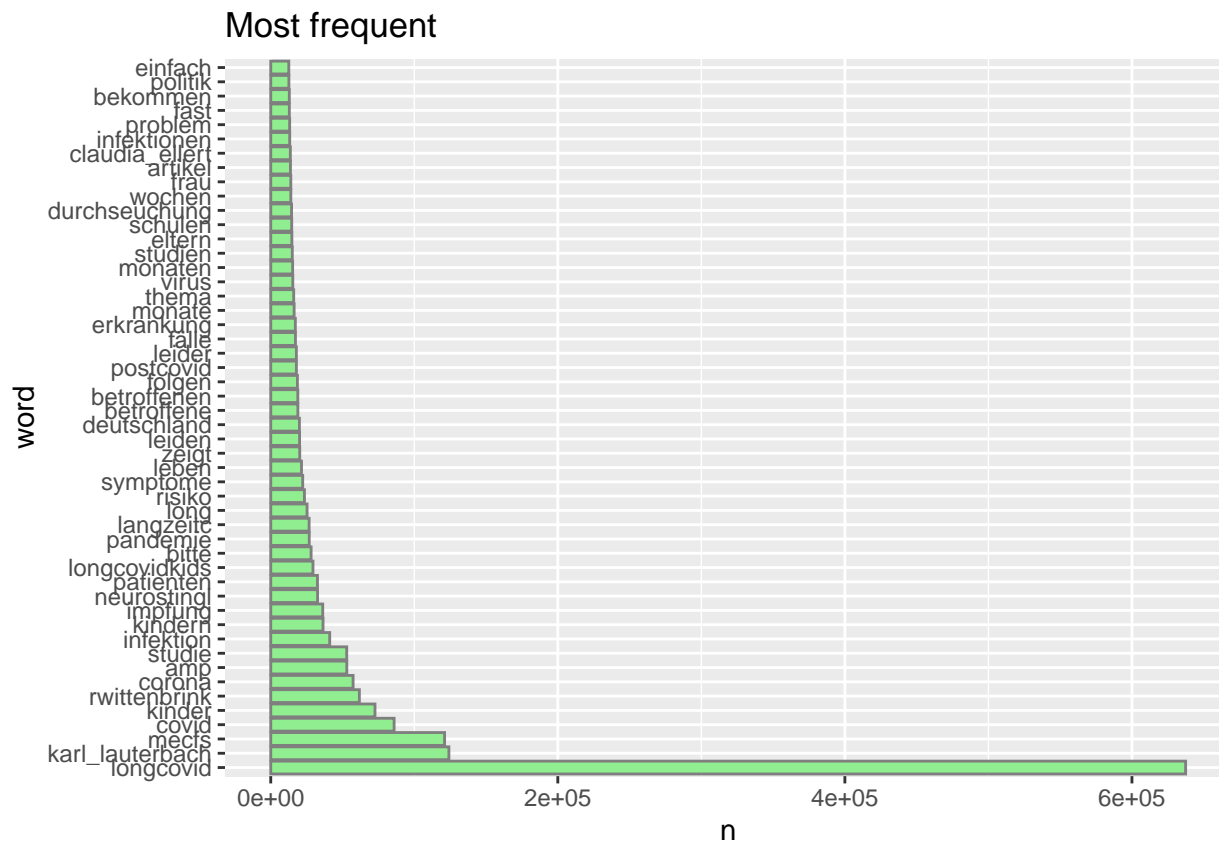
```



```

ungroup() %>%
filter(!is.na(word)) %>%
arrange(-n) %>%
top_n(50, n) %>%
mutate(word = factor(word, levels = word)) %>%
ggplot() + geom_bar(aes(word, n), stat = "identity", fill = "lightgreen",
color = "gray50") + ggtitle("Most frequent") + coord_flip()

```



Most commonly used hashtags



```
## Selecting by freq
```





```

lead_words_counts_perMonth <- lead_words %>%
  group_by(month = lubridate::floor_date(created_at, "month")) %>%
  dplyr::count(word) %>%
  ungroup() %>%
  dplyr::rename(freq = n) %>%
  arrange(-freq) %>%
  # top_n(1000)
mutate(word_percentage = freq/sum(freq))

lead_words_counts_perMonth_filtered = lead_words_counts_perMonth %>%
  filter(word %in% c("mecfs", "cfs", "fatigue", "spätfolgen"))

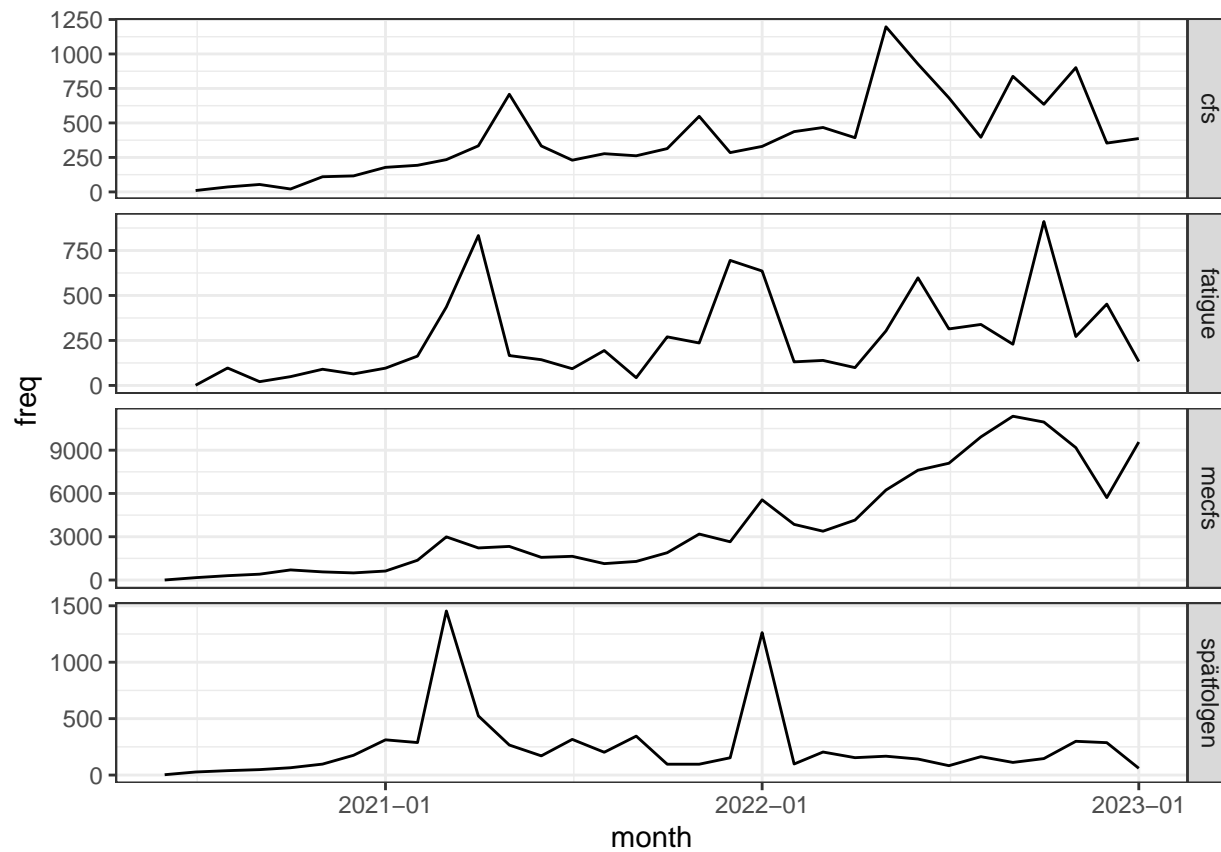
```

Plot word occurrences per month (Twitter)

```

ggplot(lead_words_counts_perMonth_filtered) + geom_line(aes(x = month,
  y = freq)) + facet_grid(rows = vars(word), scales = "free_y") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
    hjust = 1)) + scale_x_date(breaks = "1 year", date_labels = "%Y-%m") +
  theme_bw()

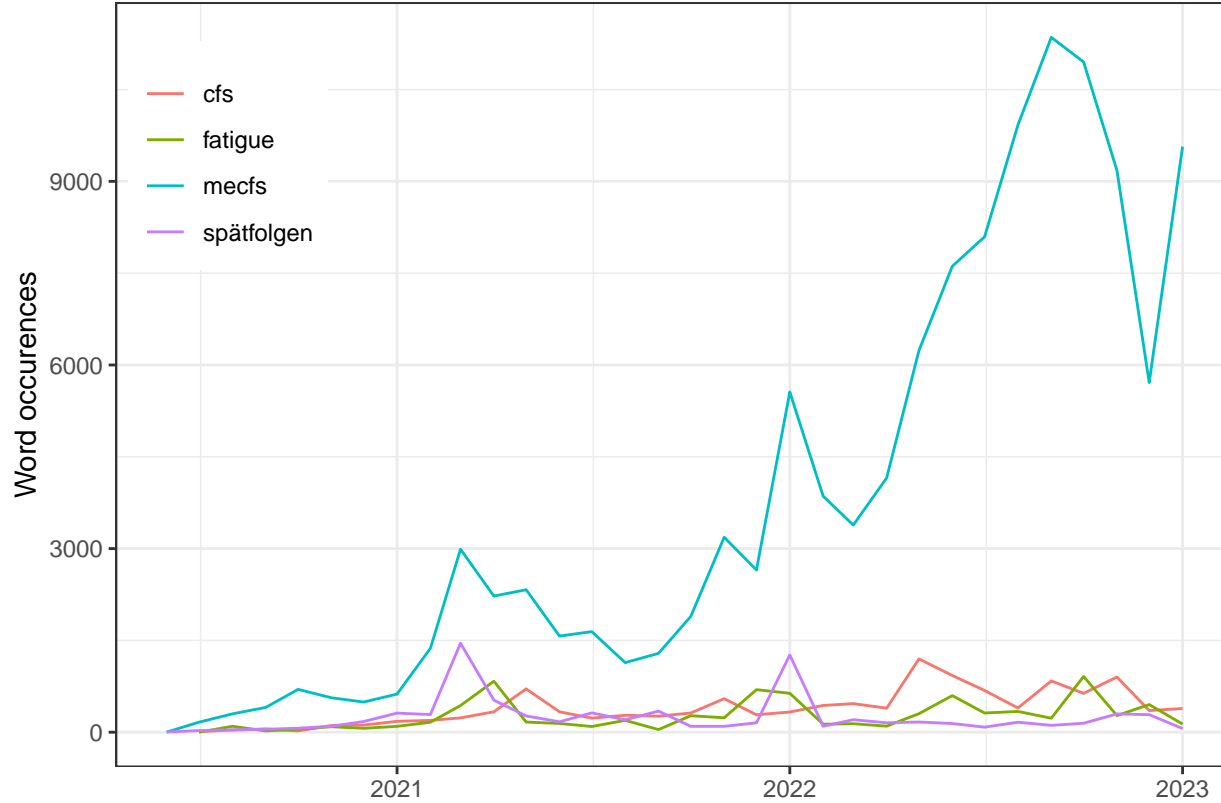
```



Compare symptome occurences

```
word_timeline <- ggplot(lead_words_counts_perMonth_filtered) +
  geom_line(aes(x = month, y = freq, color = word)) + theme_bw() +
  xlab("") + ylab("Word occurences") + theme(legend.title = element_blank(),
  legend.position = c(0.1, 0.8))
```

```
word_timeline
```



```
ggsave("timeline_words.png", word_timeline, width = 20, height = 10,
       units = "cm")
```

```
labeled_tweets <- read.table("/Users/kristinkoehler/Desktop/NLP_DL/long-covid/data/proper_dataset.csv",
                             sep = ",", header = TRUE)
```

Load and merge labelled tweets

```
tweets_merged_unique <- tweets_merged[!duplicated(tweets_merged$text),
]
```

```
labeled_merged <- merge(labeled_tweets, tweets_merged_unique,
  by.x = "uni", by.y = "text", all.y = FALSE)
```

```
head(labeled_merged)
```

```
##
```

```
## 1      „Einer der wirklich guten Tests für #LongCovid ist die endotheliale Dysfunktion mit flussvermittelter Dilatation oder
Patienten ist sie 7%“, erklärt Dr Jaco Laubscher. @gezmedinger @doctorasadkhan #DiagnoseLongCovid https://t.co/Vxri3xBLxc
```

```
## 2
```

```
## 3 @BildungslandNRW @landnrw Warum gibt es keinerlei verpflichtende Maßnahmen zum Infektionsschutz?\n\nEigenverantwortung in diesem Land
```

```
## 4                                                                                                     @Bundeskanzler @sp
```

```
## 5      @cad59 @fdpbt Großes Kino. Die @fdp hat Kindern + Eltern die Freiheit beschert, sich zwangsweise mit #Corona anzusteck
```

```
## 6      @dieMinna #Pandemie : den Corona-Jüngern läuft Corona davon, deshalb haben sie sich Hirngespinnste
```

```
## Column1 annotation_id annotator created_at.x id.x lead_time
```

```
## 1      49      378      5 2023-07-31T18:35:30.505561Z 377 10.666
```

```
## 2      78      403      5 2023-07-31T18:45:52.068846Z 402 7.646
```

```
## 3     201      519      6 2023-08-02T14:42:26.187450Z 505 11.331
```

```
## 4      67      394      5 2023-07-31T18:43:38.603013Z 393 11.605
```

```
## 5     249      518      4 2023-08-02T14:35:55.726884Z 543 21.176
```

```
## 6     219      535      6 2023-08-02T14:45:30.652132Z 521 24.974
```

```
## sentiment updated_at x id.y created_at.y
```

```
## 1 Positive 2023-07-31T18:35:30.505615Z 49 1596032602498306049 2022-11-25
```

```
## 2 Negative 2023-07-31T18:45:52.068887Z 78 1467455398647054339 2021-12-05
```

```
## 3 Negative 2023-08-02T14:42:26.187506Z 201 1554080675149529092 2022-08-01
```

```
## 4 Negative 2023-07-31T18:43:38.603048Z 67 1533681842213335041 2022-06-06
```

```
## 5 Negative 2023-08-02T14:35:55.726934Z 249 1526671305260642309 2022-05-17
```

```
## 6 Negative 2023-08-02T14:45:30.652169Z 219 1397865601306185730 2021-05-27
```

```
## author_id public_metrics.retweet_count public_metrics.reply_count
```

```
## 1      732811424      32      3
```

```
## 2 1423377255116464140      0      0
```

```
## 3 1268840416662753281      0      1
```

```
## 4 1480288705868931083      0      0
```

```
## 5 940331690383339530      0      0
```

```
## 6 1167347221135863808      0      0
```

```
## public_metrics.like_count public_metrics.quote_count
```

```
## 1      68      0
```

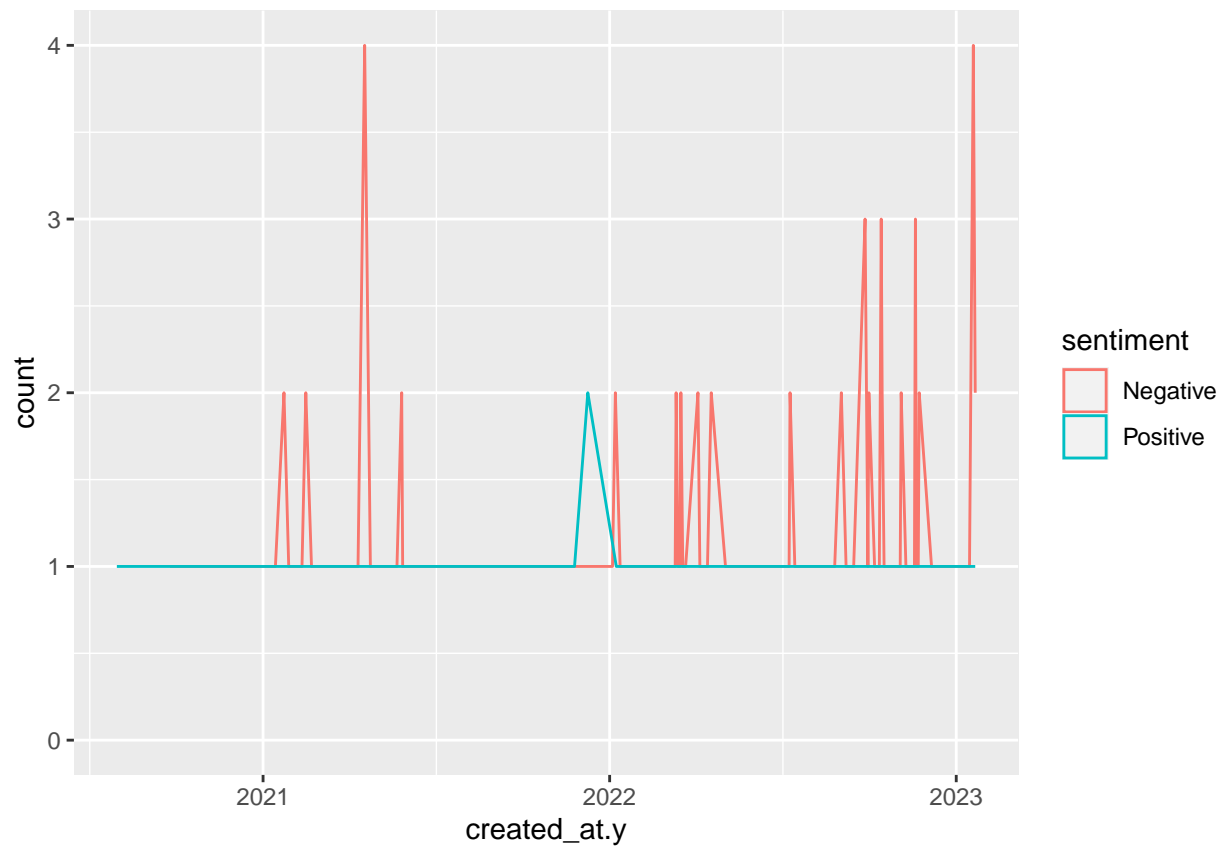
```

## 2          0          0
## 3          4          0
## 4          0          0
## 5          2          0
## 6          3          0
## public_metrics.impression_count entities.mentions.start entities.mentions.end
## 1          0          222, 235          234, 250
## 2          0          0          5
## 3          0          0, 17          16, 25
## 4          0          0, 15, 22          14, 21, 38
## 5          0          0, 7, 31          6, 13, 35
## 6          0          0          9
##          entities.mentions.username
## 1          gezmedinger , doctorasadkhan
## 2          BILD
## 3          BildungslandNRW, landnrw
## 4 Bundeskanzler , spdbt , GrueneBundestag
## 5          cad59, fdpbt, fdp
## 6          dieMinna
##          entities.mentions.id
## 1          281185181 , 2692420788
## 2          9204502
## 3          899678575204478976, 247262571
## 4 1489164711652466688, 347792540 , 24725119
## 5 25980076 , 905516414257258497, 39475170
## 6          1183228909
## entities.hashtags.start entities.hashtags.end
## 1          36, 251          46, 269
## 2          23, 54, 88, 101          34, 64, 97, 114
## 3          258, 292          277, 302
## 4          126, 134, 145, 160          133, 144, 159, 169
## 5          102, 146, 270          109, 156, 293
## 6          10, 102, 114, 193          19, 112, 128, 202
##          entities.hashtags.tag
## 1          LongCovid , DiagnoseLongCovid
## 2          Querdenker , LongCovid , dummheit , Aggressionen
## 3          Infektionsmarathon, LongCovid
## 4          Scholz , LongCovid , LongCovidKids, Familien
## 5 Corona , LongCovid , FDPrausausderRegierung

```

```
## 6      Pandemie      , LongCovid      , LongCovidKids, Mutanten
##
## 1      270, 293, https://t.co/Vxri3xBLxc, https://twitter.com/loscharlos/status/1595910352709251072, twitter.com/loscharlos/sta...
## 2
## 3
## 4 170, 193, https://t.co/Ord4rbexbV, https://twitter.com/AugersMonster/status/1533599169331359744, twitter.com/AugersMonster/...
## 5
## 6
##      year
## 1 2022
## 2 2021
## 3 2022
## 4 2022
## 5 2022
## 6 2021
```

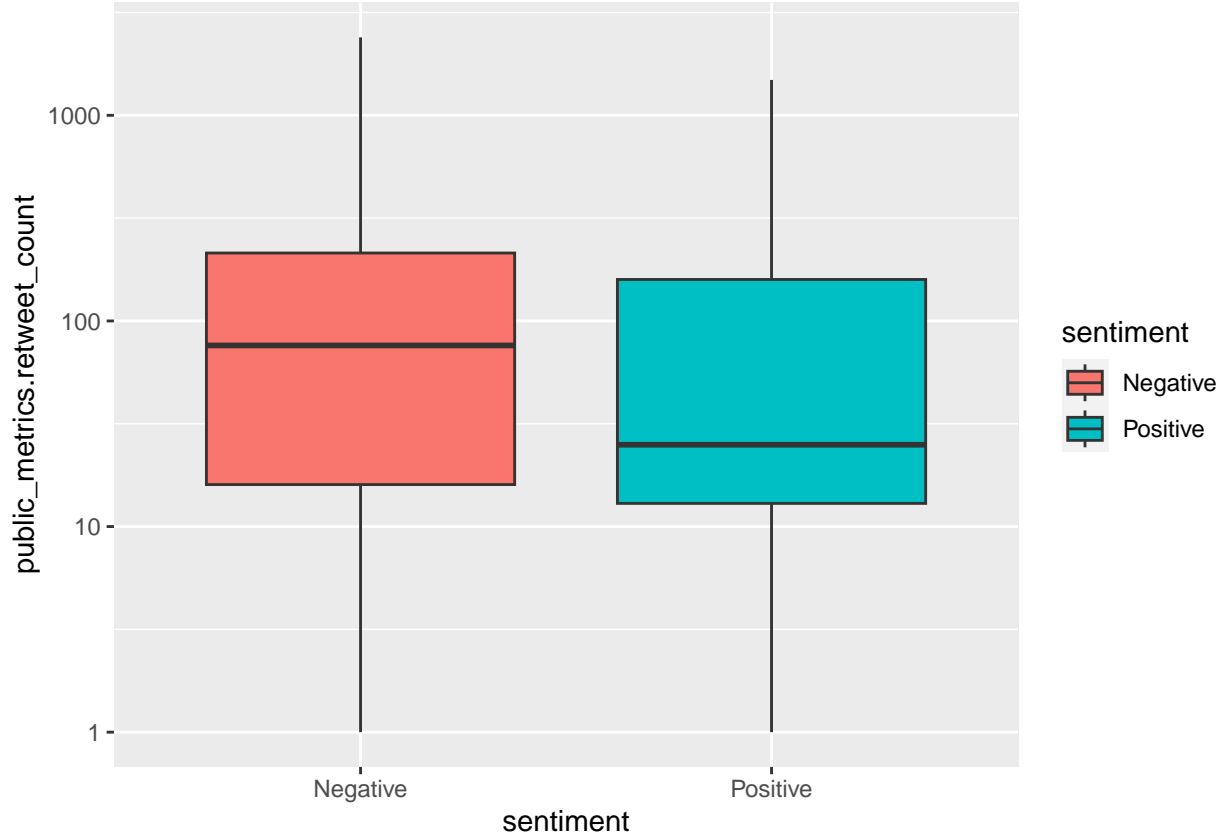
```
labeled_merged$created_at.y <- as.Date(labeled_merged$created_at.y,
  format = "%Y-%m-%d")
ggplot(labeled_merged) + geom_density(aes(x = created_at.y, color = sentiment),
  stat = "count")
```



```
ggplot(labeled_merged) + geom_boxplot(aes(x = sentiment, y = public_metrics.retweet_count,
  fill = sentiment)) + scale_y_continuous(trans = "log10")
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 17 rows containing non-finite values (`stat_boxplot()`).
```



```
ggplot(labeled_merged) + geom_boxplot(aes(x = sentiment, y = public_metrics.like_count,  
  fill = sentiment)) + scale_y_continuous(trans = "log10")
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 220 rows containing non-finite values (`stat_boxplot()`).
```



