Textmining Becker

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
library(tidyr)
library(tidytext)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
#usa pro Monat
daten_usa<-read.csv("C:/Users/Christian/Documents/textmining/R-projekt/BeckerSeminar2/Testing/Daten2012u
data_fr_usa<- data.frame(daten_usa, stringsAsFactors=FALSE)
data_fr_usa$Tweets<-as.character(data_fr_usa$Tweets)
tidy_daten2012_word <- data_fr_usa %>% unnest_tokens(word, Tweets)
#entferne stopwords
tidy_2012_ohne_stopwords <- tidy_daten2012_word %>% anti_join(stop_words)
```

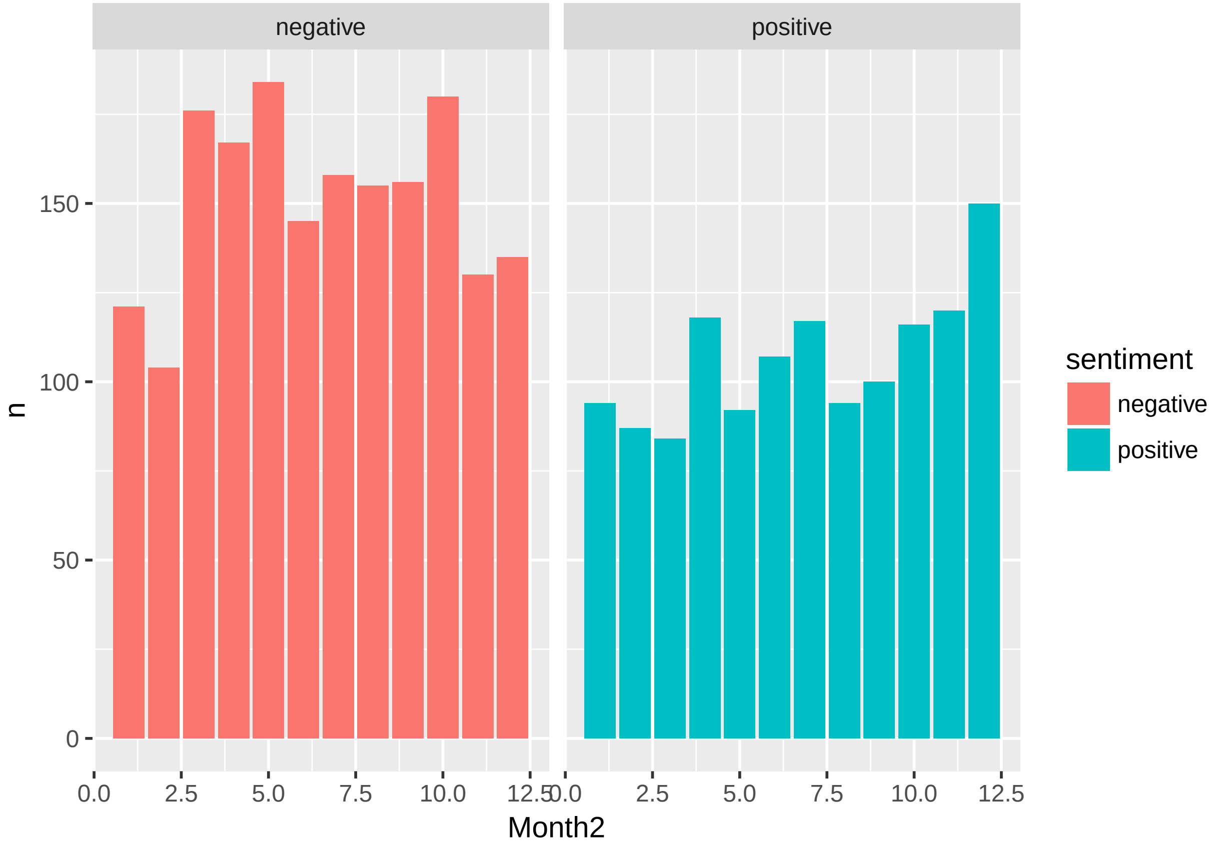
```
## Joining, by = "word"
```

```
#join bing
tidy_2012_ohne_stopwords$Month2<-NULL
tidy_2012_ohne_stopwords[tidy_2012_ohne_stopwords$Month=="Jan","Month2"]<- month(01)
tidy_2012_ohne_stopwords[tidy_2012_ohne_stopwords$Month=="Feb","Month2"]<- month(02)
tidy_2012_ohne_stopwords[tidy_2012_ohne_stopwords$Month=="Mar","Month2"]<- month(03)
tidy_2012_ohne_stopwords[tidy_2012_ohne_stopwords$Month=="Apr","Month2"]<- month(04)
```

```
tidy_2012_ohne_stopwords[tidy_2012_ohne_stopwords$Month=="May","Month2"]<- month(05)
tidy_2012_ohne_stopwords[tidy_2012_ohne_stopwords$Month=="Jun","Month2"]<- month(06)
tidy_2012_ohne_stopwords[tidy_2012_ohne_stopwords$Month=="Jul","Month2"]<- month(07)
tidy_2012_ohne_stopwords[tidy_2012_ohne_stopwords$Month=="Aug","Month2"]<- month(08)
tidy_2012_ohne_stopwords[tidy_2012_ohne_stopwords$Month=="Sep","Month2"]<- month(09)
tidy_2012_ohne_stopwords[tidy_2012_ohne_stopwords$Month=="Oct","Month2"]<- month(10)
tidy_2012_ohne_stopwords[tidy_2012_ohne_stopwords$Month=="Nov","Month2"]<- month(11)
tidy_2012_ohne_stopwords[tidy_2012_ohne_stopwords$Month=="Dec","Month2"]<- month(12)
bing <- get_sentiments("bing")
datplot<-tidy_2012_ohne_stopwords  %>%
  inner_join(bing) %>%
  group_by(Month2)%>%
  count(sentiment)
```

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

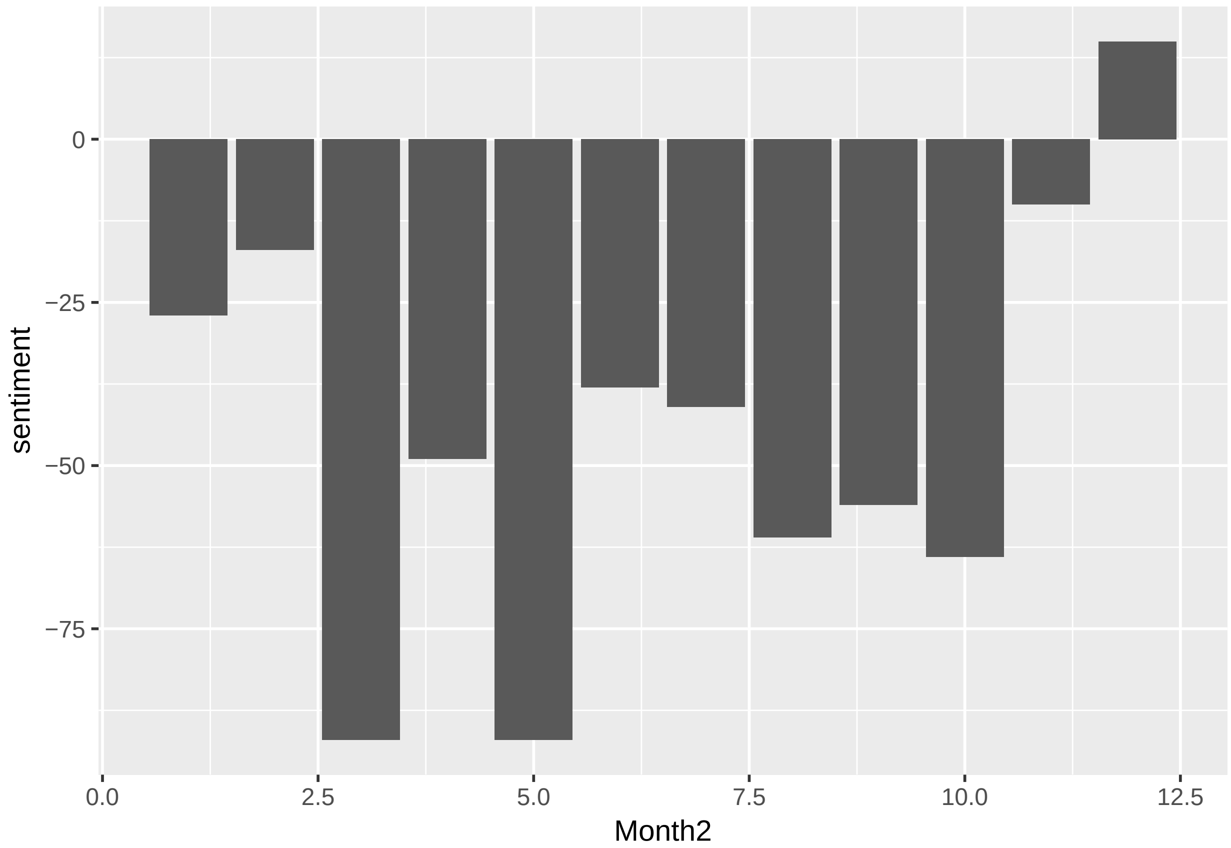
```
ggplot(data=datplot, aes(x=Month2, y=n, fill=sentiment)) + geom_col(show.legend = FALSE)+
  geom_bar( aes(x=Month2, y=n),stat="identity") + facet_wrap(~sentiment, ncol = 2, scales = "free_x")
```



```
#differrenz
dif_us<-tidy_2012_ohne_stopwords  %>%
  inner_join(bing) %>%
  group_by(Month2)%>%
  count(sentiment) %>%
  spread(sentiment, n)%>%
  mutate(sentiment = positive - negative)
```

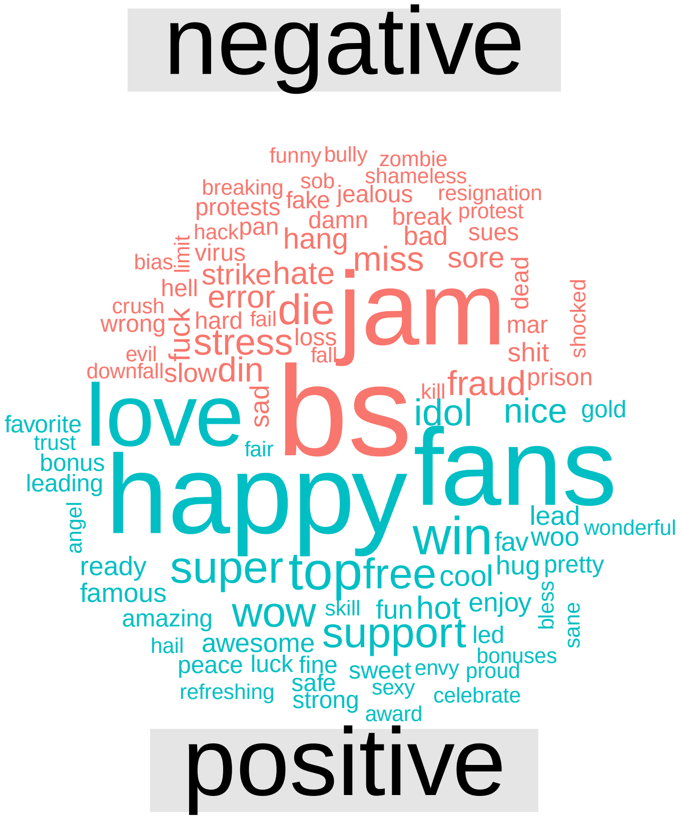
```
## Joining, by = "word"
```

```
ggplot(data=dif_us, aes(x=Month2, y=sentiment),fill=sentiment) + geom_col(show.legend = FALSE)+
  geom_bar(stat="identity")
```



```
#wordcloud usa------------------------------------------
word_cloud_usa<-tidy_2012_ohne_stopwords  %>%
  inner_join(bing) %>%
count(word, sentiment, sort = TRUE) %>%
  acast(word ~ sentiment, value.var = "n", fill = 0) %>%
  comparison.cloud(colors = c("#F8766D", "#00BFC4"),
                   max.words = 100)
```

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

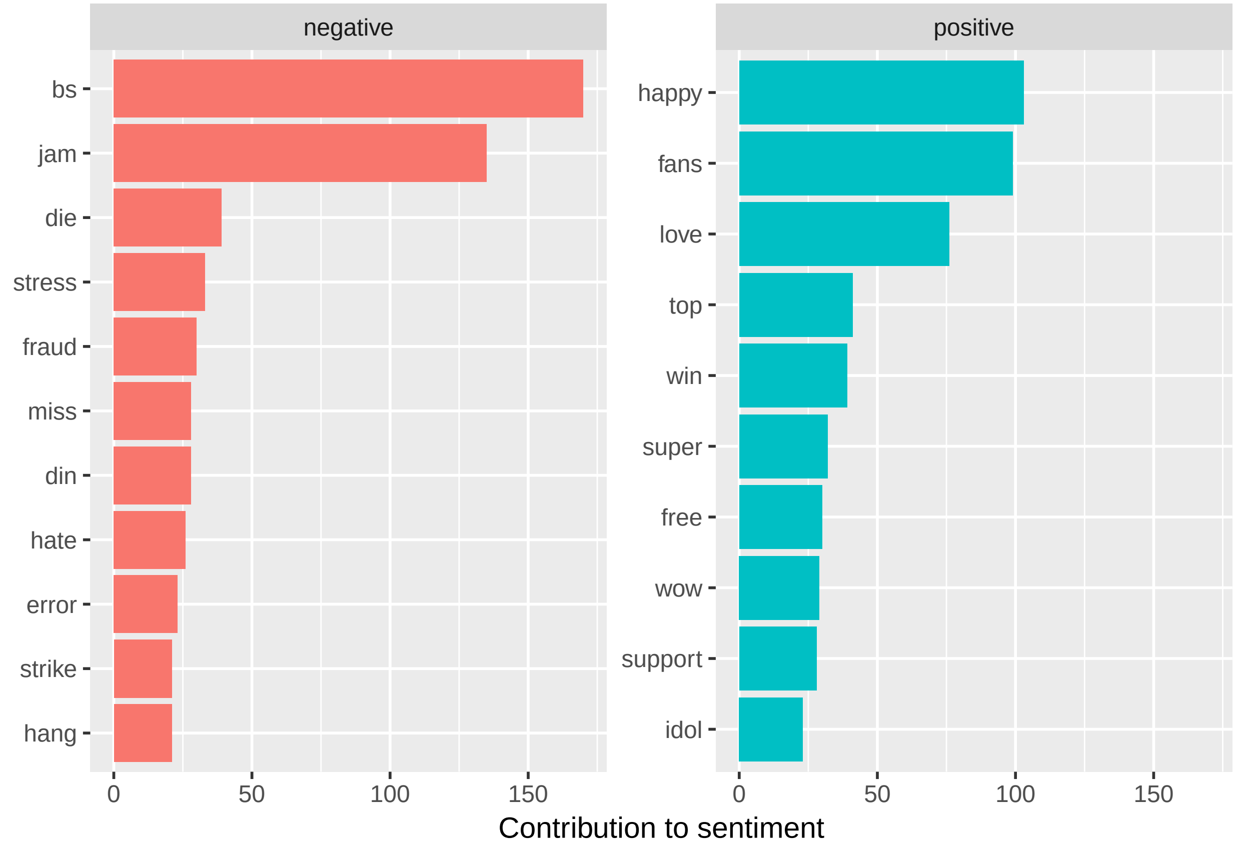


```
#Die Ranking postiv und negative Wörts USA-------------------------------------------------------------
wordcount <-tidy_2012_ohne_stopwords %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup()
```

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

```
wordcount %>%
  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()
```

```
## Selecting by n
```



```
#Vergleich Nrc, Bing und AFINN USA---------------------------------------------------------------------

afinn <- tidy_2012_ohne_stopwords%>%
  inner_join(get_sentiments("afinn")) %>%
  group_by(Month2) %>%
  summarise(sentiment = sum(score)) %>%
  mutate(method = "AFINN")
```

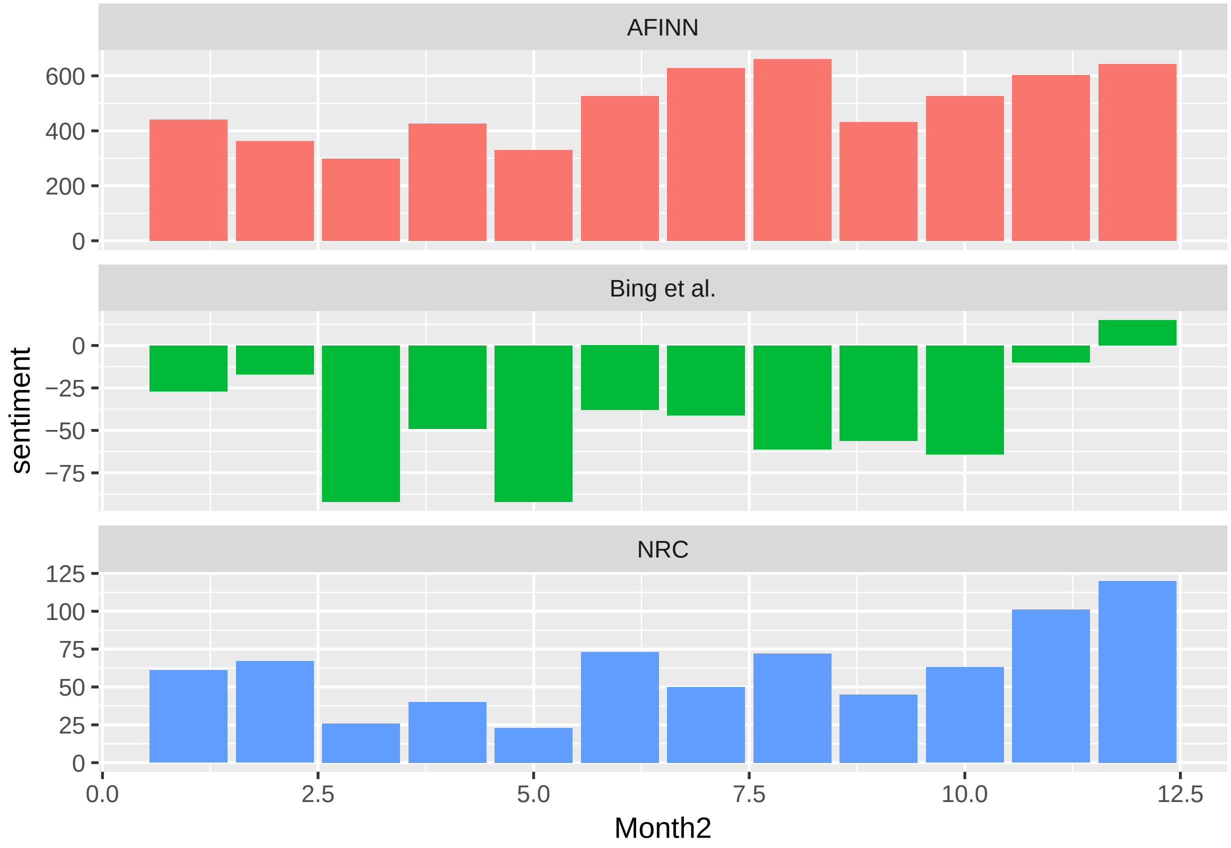
```
## Joining, by = "word"
```

```
bing_and_nrc <- bind_rows(tidy_2012_ohne_stopwords%>%
                            inner_join(get_sentiments("bing")) %>%
                            mutate(method = "Bing et al."),
                          tidy_2012_ohne_stopwords %>%
                            inner_join(get_sentiments("nrc")) %>%
                                          filter(sentiment %in% c("positive","negative"))) %>%
                            mutate(method = "NRC")) %>%
  count(method, Month2, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative)
```

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

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

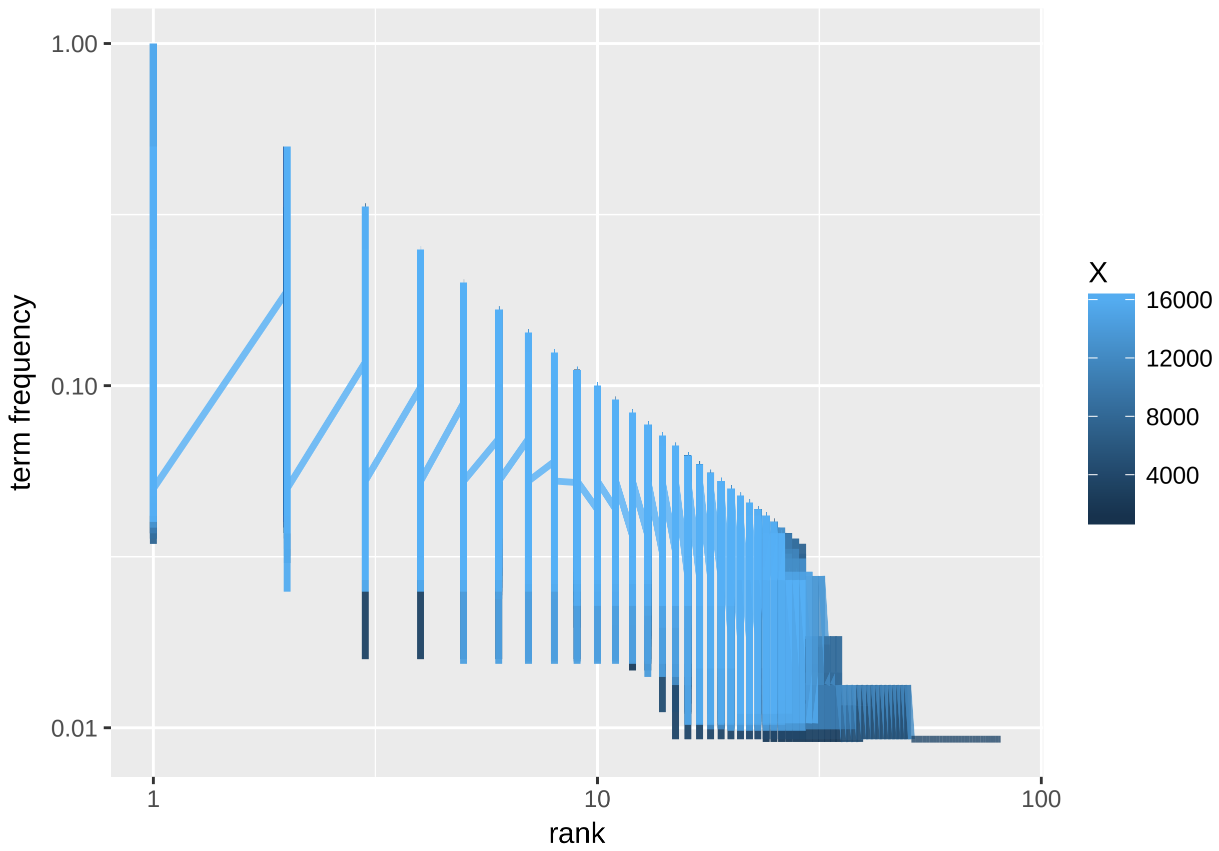
```
bind_rows(afinn,
          bing_and_nrc) %>%
  ggplot(aes(Month2, sentiment, fill = method)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~method, ncol = 1, scales = "free_y")
```



```
#frequency id-tf usa----------------------------------------------------------------------------
tweets_words <- tidy_2012_ohne_stopwords %>%
  count(X, word, sort = TRUE) %>%
  ungroup()

total_words <- tweets_words %>%
  group_by(X) %>%
  summarize(total = sum(n))

tweets_words <- left_join(tweets_words, total_words)
```

```
## Joining, by = "X"
```

```
freq_by_rank <- tweets_words %>%
  group_by(X) %>%
  mutate(rank = row_number(),
         `term frequency` = n/total)

freq_by_rank %>%
  ggplot(aes(rank, `term frequency`, color = X)) +
  geom_line(size = 1.2, alpha = 0.8) +
  scale_x_log10() +
  scale_y_log10()
```



```
#n-gramme-usa---------------------------------------------------

#Griechenland pro Monat----------------------------------------------------------------
daten_griechenland<-read.csv("C:/Users/Christian/Documents/textmining/R-projekt/BeckerSeminar2/Testing/
data_fr_griechenland<- data.frame(daten_griechenland)
data_fr_griechenland$Tweets<-as.character(data_fr_griechenland$Tweets)
tidy_daten2012_word_gr <- data_fr_griechenland %>% unnest_tokens(word, Tweets)
#entferne stopwords
tidy_2012_ohne_stopwords_gr <- tidy_daten2012_word_gr %>% anti_join(stop_words)
```

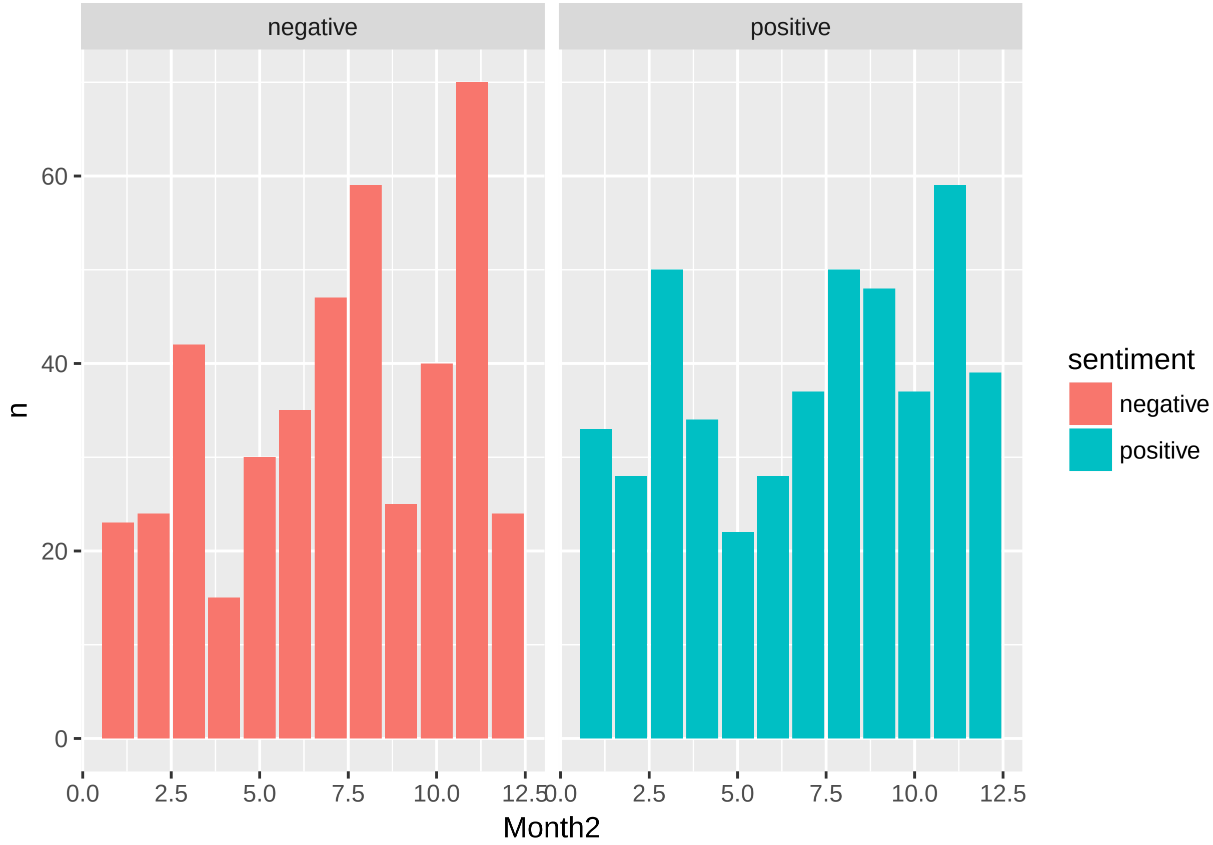
```
## Joining, by = "word"
```

```
#join bing
tidy_2012_ohne_stopwords$Month2<-NULL
tidy_2012_ohne_stopwords_gr[tidy_2012_ohne_stopwords_gr$Month=="Jan","Month2"]<- month(01)
tidy_2012_ohne_stopwords_gr[tidy_2012_ohne_stopwords_gr$Month=="Feb","Month2"]<- month(02)
tidy_2012_ohne_stopwords_gr[tidy_2012_ohne_stopwords_gr$Month=="Mar","Month2"]<- month(03)
tidy_2012_ohne_stopwords_gr[tidy_2012_ohne_stopwords_gr$Month=="Apr","Month2"]<- month(04)
tidy_2012_ohne_stopwords_gr[tidy_2012_ohne_stopwords_gr$Month=="May","Month2"]<- month(05)
tidy_2012_ohne_stopwords_gr[tidy_2012_ohne_stopwords_gr$Month=="Jun","Month2"]<- month(06)
tidy_2012_ohne_stopwords_gr[tidy_2012_ohne_stopwords_gr$Month=="Jul","Month2"]<- month(07)
tidy_2012_ohne_stopwords_gr[tidy_2012_ohne_stopwords_gr$Month=="Aug","Month2"]<- month(08)
tidy_2012_ohne_stopwords_gr[tidy_2012_ohne_stopwords_gr$Month=="Sep","Month2"]<- month(09)
tidy_2012_ohne_stopwords_gr[tidy_2012_ohne_stopwords_gr$Month=="Oct","Month2"]<- month(10)
tidy_2012_ohne_stopwords_gr[tidy_2012_ohne_stopwords_gr$Month=="Nov","Month2"]<- month(11)
tidy_2012_ohne_stopwords_gr[tidy_2012_ohne_stopwords_gr$Month=="Dec","Month2"]<- month(12)
bing <- get_sentiments("bing")
```

```
datplot_gr<-tidy_2012_ohne_stopwords_gr  %>%
  inner_join(bing) %>%
  group_by(Month2)%>%
  count(sentiment)
```

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

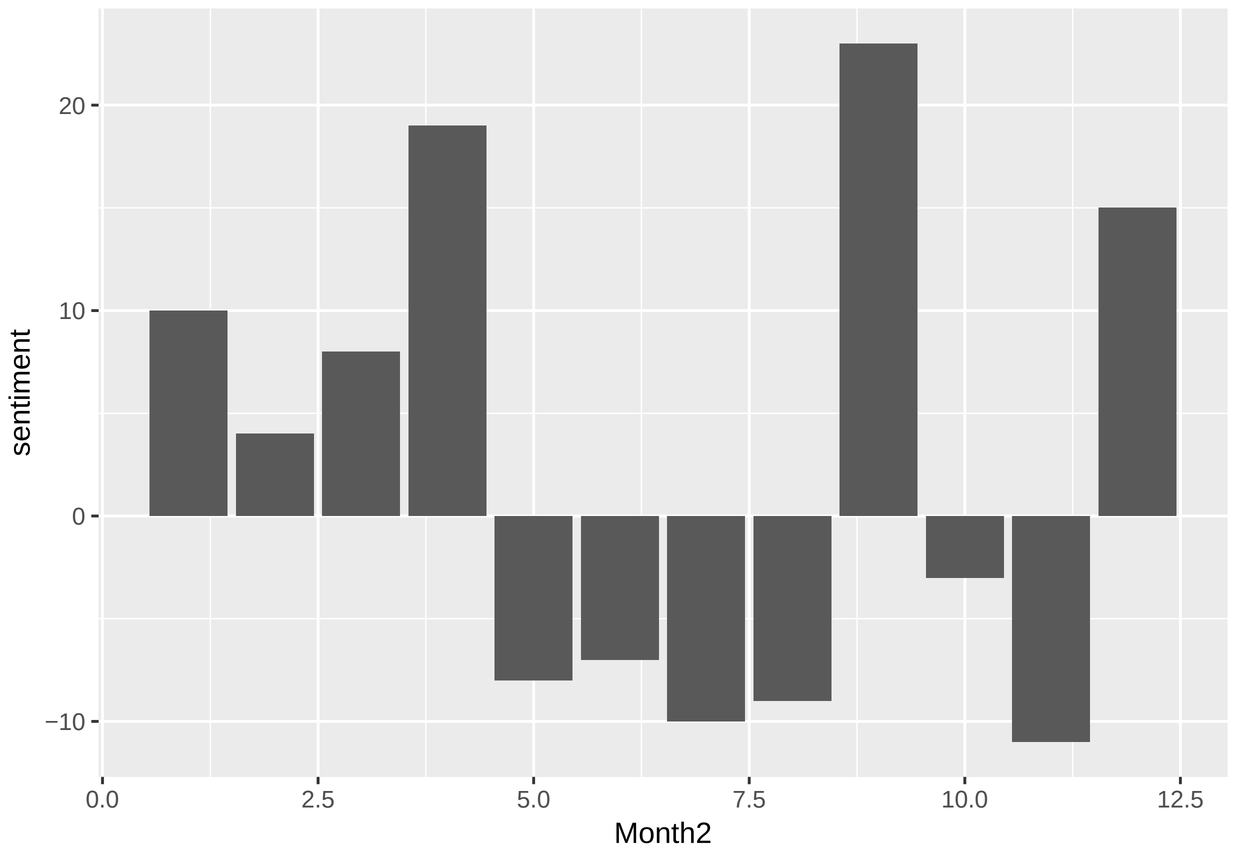
```
ggplot(data=datplot_gr, aes(x=Month2, y=n, fill=sentiment)) + geom_col(show.legend = FALSE)+
  geom_bar(stat="identity") + facet_wrap(~sentiment, ncol = 2, scales = "free_x")
```



```
#differrenz-----------------------------------------------------------------------------------
dif_griechen<-tidy_2012_ohne_stopwords_gr  %>%
  inner_join(bing) %>%
  group_by(Month2)%>%
  count(sentiment) %>%
  spread(sentiment, n)%>%
  mutate(sentiment = positive - negative)
```

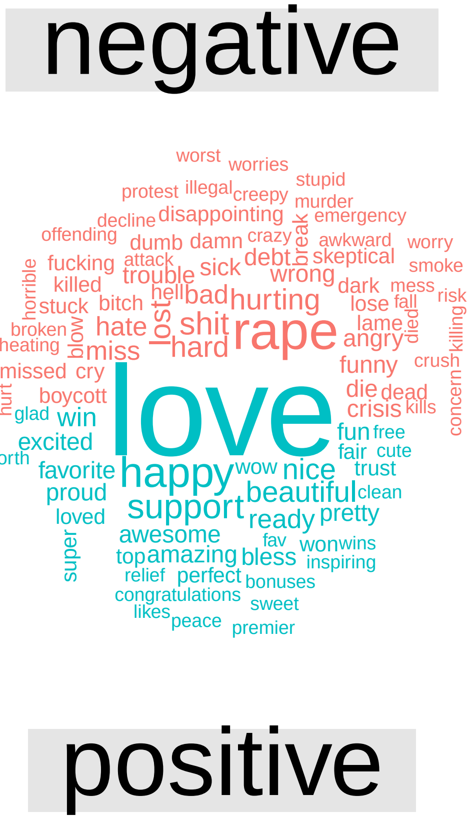
```
## Joining, by = "word"
```

```
ggplot(data=dif_griechen, aes(x=Month2, y=sentiment),fill=sentiment) + geom_col(show.legend = FALSE)+
  geom_bar(stat="identity")
```



```
#wordcloud griechenland-----------------------------------------
word_cloud_usa<-tidy_2012_ohne_stopwords_gr  %>%
  inner_join(bing) %>%
  count(word, sentiment, sort = TRUE) %>%
  acast(word ~ sentiment, value.var = "n", fill = 0) %>%
  comparison.cloud(colors = c("#F8766D", "#00BFC4"),
                   max.words = 100)
```

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

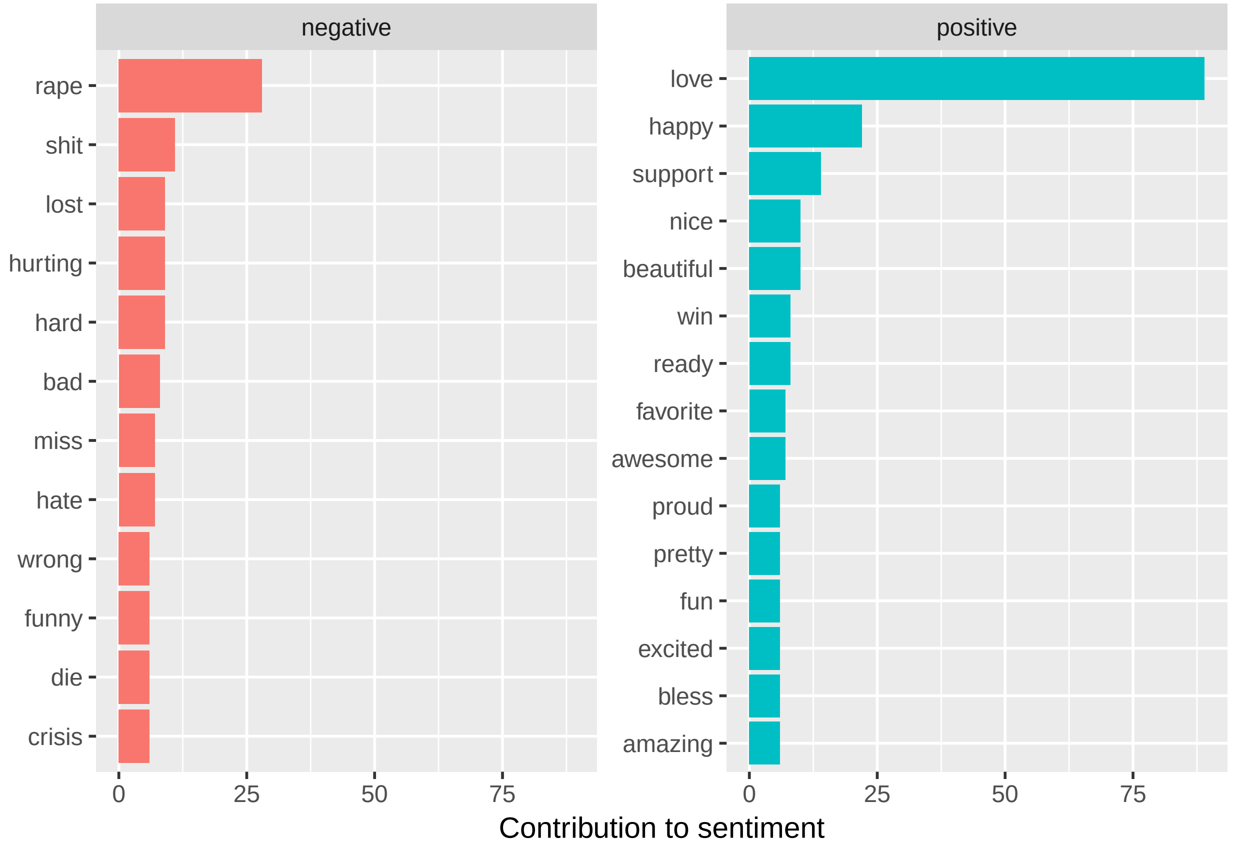


```r
#Die Ranking postiv und negative Wörts Griechenland--------------------------------------------------
wordcount <-tidy_2012_ohne_stopwords_gr %>%
inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup()
```

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

```r
wordcount %>%
  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()
```

```
## Selecting by n
```



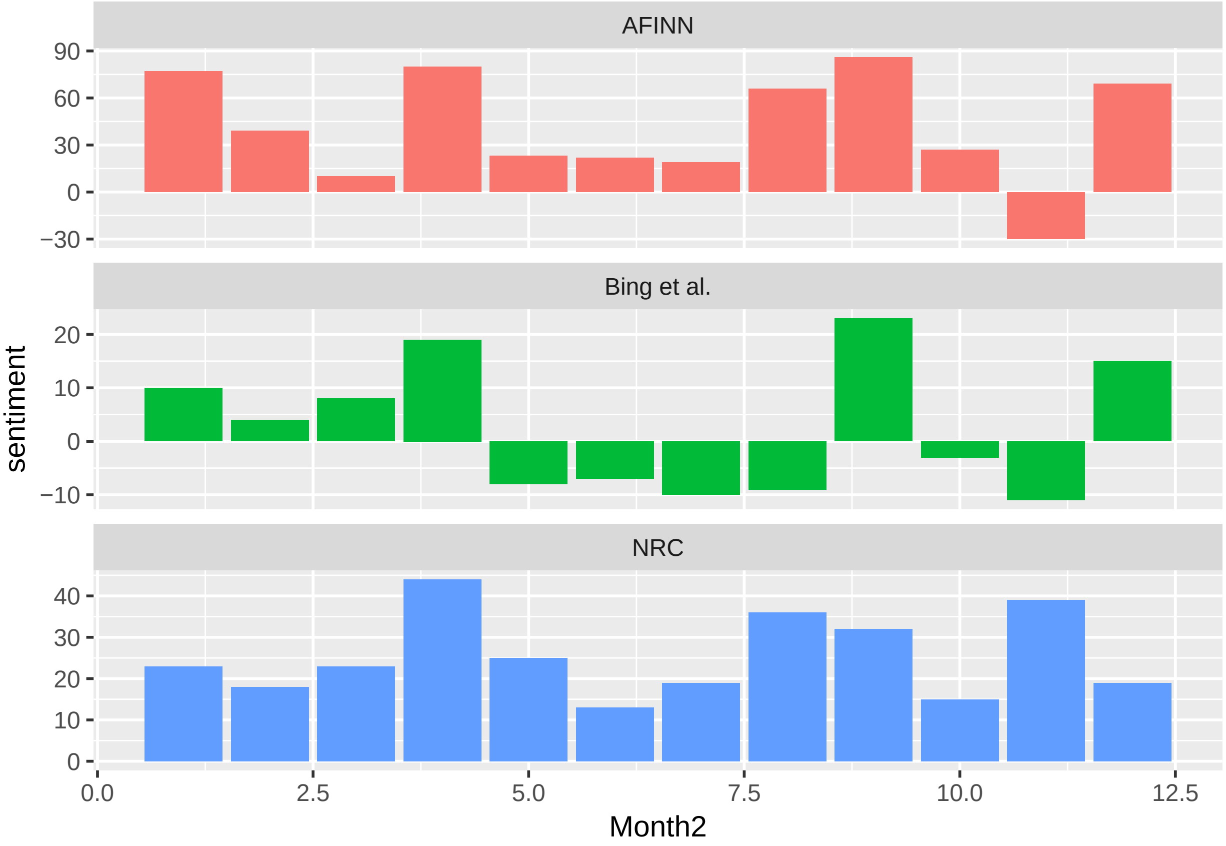
```
#Vergleich Nrc, Bing und AFINN Griechenland------------------------------------------------------------

afinn_gr <- tidy_2012_ohne_stopwords_gr%>%
  inner_join(get_sentiments("afinn")) %>%
  group_by(Month2) %>%
  summarise(sentiment = sum(score)) %>%
  mutate(method = "AFINN")
```

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

```
bing_and_nrc_gr <- bind_rows(tidy_2012_ohne_stopwords_gr%>%
                            inner_join(get_sentiments("bing")) %>%
                            mutate(method = "Bing et al."),
                          tidy_2012_ohne_stopwords_gr %>%
                            inner_join(get_sentiments("nrc")) %>%
                                         filter(sentiment %in% c("positive","negative"))) %>%
                            mutate(method = "NRC")) %>%
  count(method, Month2, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative)
```

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

```
bind_rows(afinn_gr,
          bing_and_nrc_gr) %>%
  ggplot(aes(Month2, sentiment, fill = method)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~method, ncol = 1, scales = "free_y")
```



```
class(tidy_daten2012_word$Month)
```

```
## [1] "factor"
```