Mini Project Part 3

Nick Starcevich

Introduction

In the past, I've heard people complain about how the news is always too negative and emotional to listen to or read and that it never has any happy stories. I have heard of this data set containing New York Times article titles in the 90s and 2000s. I wanted to see if this seem to be true with these titles, so I did some analysis on this data set to see whether this may be true or not.

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(RTextTools)
```

```
Warning: package 'RTextTools' was built under R version 4.3.3
```

```
Loading required package: SparseM
```

```
Attaching package: 'SparseM'
```

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

    backsolve
```

```
library(tidytext)
```

```
Warning: package 'tidytext' was built under R version 4.3.3
```

```
data(NYTimes)
NYtimes <- NYTimes |>
  as_tibble() |>
  mutate(Date = dmy(Date))

#explain why looking at upper case words in particular (upper case means bigger deal and u
UppercaseNYT <- NYtimes |>
  mutate(Title = as.character(Title)) |>
  filter(str_detect(Title, "\\b[A-Z]+\\b"))


UppercaseNYT_clean <- UppercaseNYT |>
  unnest_tokens(word, Title)
```

In the past, I've noticed that people tend to express more emotion using words containing all uppercase letters. In this analysis I compare sentiments in titles with lowercase letters to titles with uppercase letters, and I also compare those uppercase titles to subtitles within those titles to see if they are any different.

```
Normal_title_NYT <- NYtimes |>
  mutate(Title = as.character(Title)) |>
  anti_join(UppercaseNYT, by = c("Title" = "Title"))

Normal_title_NYT_clean <- Normal_title_NYT |>
  unnest_tokens(word, Title)

NYT_sentiments_values_lower <- Normal_title_NYT_clean |>
  inner_join(get_sentiments("afinn"))
```

```
Joining with `by = join_by(word)`
```

```
  NYT_sentiments_values_lower
```

```
# A tibble: 1,208 x 6
   Article_ID Date       Subject                           Topic.Code word  value
        <int> <date>     <fct>                                  <int> <chr> <dbl>
 1      41246 1996-01-01 Jails overwhelmed with hardened~          12 stru~    -2
 2      41268 1996-01-03 Contenders for 1996 Presedentia~          20 cost~    -2
 3      41279 1996-01-04 Bosnian Serb leader criticized ~          19 top       2
 4      41279 1996-01-04 Bosnian Serb leader criticized ~          19 atta~    -1
 5      41314 1996-01-08 economists not afraid of a defi~           1 fear     -2
 6      41314 1996-01-08 economists not afraid of a defi~           1 defi~    -2
 7      41344 1996-01-11 census changes                            20 cool      1
 8      41418 1996-01-19 uneasy time in saudi arabia               19 unea~    -2
 9      41418 1996-01-19 uneasy time in saudi arabia               19 hope      2
10      41439 1996-01-21 space shuttle lands                       17 grea~     3
# i 1,198 more rows
```

```
  NYT_sentiments_values_upper <- UppercaseNYT_clean |>
    inner_join(get_sentiments("afinn"))
```

```
Joining with `by = join_by(word)`
```

```
  NYT_sentiments_values_upper
```

```
# A tibble: 1,084 x 6
   Article_ID Date       Subject                           Topic.Code word  value
        <int> <date>     <fct>                                  <int> <chr> <dbl>
 1      41290 1996-01-05 Battle over budget: Republican ~           1 batt~    -1
 2      41290 1996-01-05 Battle over budget: Republican ~           1 drop     -1
 3      41333 1996-01-10 budget fight                               1 batt~    -1
 4      41355 1996-01-12 barneys seeks bankruptcy                  15 fight    -1
 5      41367 1996-01-14 clinton in bosnia                         19 thank     2
 6      41379 1996-01-15 AIDS HMOs                                  3 care      2
 7      41379 1996-01-15 AIDS HMOs                                  3 trou~    -2
 8      41496 1996-01-28 murder suspect arrested                   12 dream     1
 9      41518 1996-01-30 new york governor and crime               24 crime    -3
10      41518 1996-01-30 new york governor and crime               24 susp~    -1
# i 1,074 more rows
```

```r
PostColon <- UppercaseNYT |>
  mutate(TitlesPostColon = str_extract(Title, ":.*")) |>
  select(-Title) |>
  drop_na() |>
  mutate(TitlesPostColon = str_extract(TitlesPostColon, "[A-Z].*"))
PostColon
```

```
# A tibble: 432 x 5
   Article_ID Date       Subject                       Topic.Code TitlesPostColon
        <int> <date>     <fct>                              <int> <chr>
 1      41290 1996-01-05 Battle over budget: Republi~           1 THE OVERVIEW; ~
 2      41333 1996-01-10 budget fight                            1 THE OVERVIEW; ~
 3      41355 1996-01-12 barneys seeks bankruptcy               15 THE DIFFICULTI~
 4      41428 1996-01-20 Democrat Representative dis~           20 IN THE SOUTH; ~
 5      41621 1996-02-11 marriage in japan                      19 For Better or ~
 6      41758 1996-02-26 republican primary                     20 THE PRIMARY SP~
 7      41770 1996-02-27 bob dole campaign shakeup              20 BOB DOLE; With~
 8      41792 1996-02-29 south carolina primary                 20 THE ISSUES; So~
 9      41815 1996-03-03 dole wins in south carolina            20 CHANGING DIREC~
10      41883 1996-03-11 pat buchanan campaign                  20 PATRICK J. BUC~
# i 422 more rows
```

```r
PostColon_clean <- PostColon |>
  unnest_tokens(word, TitlesPostColon)

NYT_sentiments_values_post_colon <- PostColon_clean |>
  inner_join(get_sentiments("afinn"))
```

```
Joining with `by = join_by(word)`
```

```r
NYT_sentiments_values_post_colon
```

```
# A tibble: 130 x 6
   Article_ID Date       Subject                           Topic.Code word  value
        <int> <date>     <fct>                                  <int> <chr> <dbl>
 1      41290 1996-01-05 Battle over budget: Republican ~           1 drop     -1
 2      41355 1996-01-12 barneys seeks bankruptcy                  15 fight    -1
 3      41621 1996-02-11 marriage in japan                         19 bett~     2
 4      41621 1996-02-11 marriage in japan                         19 worse    -3
```

```
 5      41621 1996-02-11 marriage in japan                    19 love      3
 6      41815 1996-03-03 dole wins in south carolina          20 win       4
 7      41883 1996-03-11 pat buchanan campaign                20 fun       4
 8      42085 1996-04-01 dick armey                           20 no       -1
 9      42108 1996-04-04 crash kills commerce secretary       30 resc~     2
10      42119 1996-04-05 investigation into crash             30 bad      -3
# i 120 more rows
```

Analysis

```
NYT_sentiments_values_lower|>
  summarize(average_sentiment_value_lower = mean(value))
```

```
# A tibble: 1 x 1
  average_sentiment_value_lower
                          <dbl>
1                        -0.690
```

```
NYT_sentiments_values_upper|>
  summarize(average_sentiment_value_caps = mean(value))
```

```
# A tibble: 1 x 1
  average_sentiment_value_caps
                         <dbl>
1                       -0.822
```

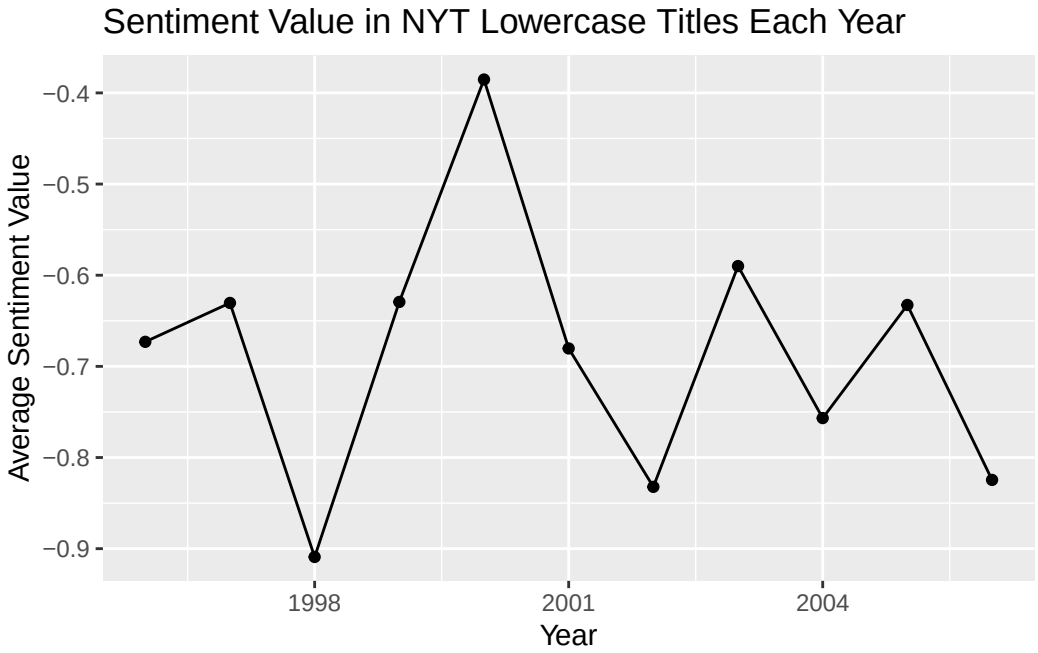
```
NYT_sentiments_values_post_colon|>
  summarize(average_sentiment_value_caps = mean(value))
```

```
# A tibble: 1 x 1
  average_sentiment_value_caps
                         <dbl>
1                       -0.569
```

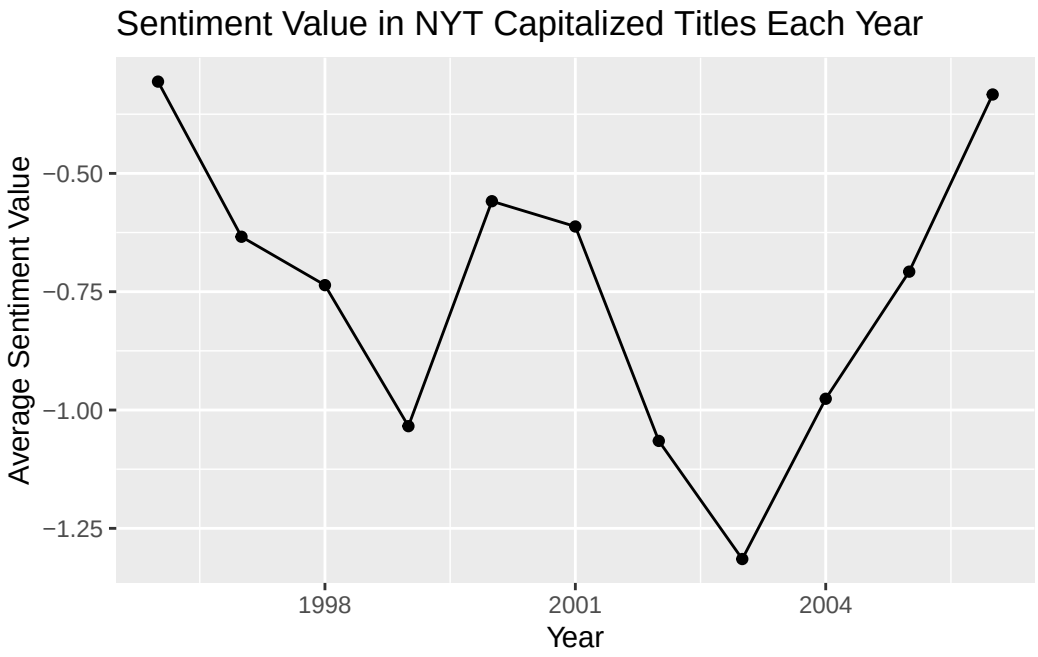
The sentimental value scale works as following: the higher the distance away from zero the values are, the more sentimental it is, and if the value is positive, the sentiments are positive sentiments while if the values are negative, they are negative sentiments. Now we can see that in all three cases that we have an average of slightly negative words being used in these titles

whereas predicted the most sentimental titles are all upper ones. Even though the subtitles are also in all uppercase the sentiments are slightly more positive than the lowercase data.

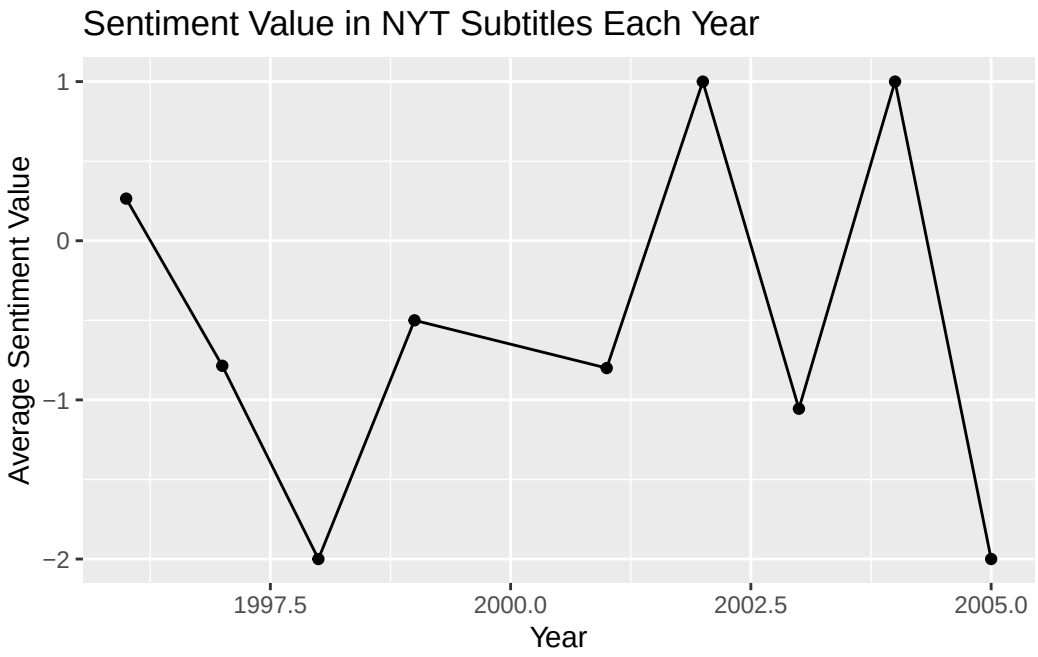
```
NYT_sentiments_values_lower |>
  mutate(year = str_sub(Date, 1, 4)) |>
  group_by(year) |>
  summarize(average_sentiment_value_lower = mean(value)) |>
  ggplot(aes(x = as.numeric(year), y = average_sentiment_value_lower)) +
  geom_point() +
  geom_line() +
  labs(title = "Sentiment Value in NYT Lowercase Titles Each Year", x = "Year", y = "Avera
```




```
NYT_sentiments_values_upper |>
  mutate(year = str_sub(Date, 1, 4)) |>
  group_by(year) |>
  summarize(average_sentiment_value_caps = mean(value)) |>
  ggplot(aes(x = as.numeric(year), y = average_sentiment_value_caps)) +
  geom_point() +
  geom_line() +
  labs(title = "Sentiment Value in NYT Capitalized Titles Each Year", x = "Year", y = "Ave
```




```
NYT_sentiments_values_post_colon |>
  mutate(year = str_sub(Date, 1, 4)) |>
  group_by(year) |>
  summarize(average_sentiment_value_caps = mean(value)) |>
  ggplot(aes(x = as.numeric(year), y = average_sentiment_value_caps)) +
  geom_point() +
  geom_line() +
  labs(title = "Sentiment Value in NYT Subtitles Each Year", x = "Year", y = "Average Sent
```



Here we compare the sentimental value for each of our three categories by year. There aren't any trends that are notable between the three that are similar. However, it is interesting to see if you look at the scales, that all uppercase titles are clearly more negatively sentimental than the others. It is also interesting to see that there are some average year values in the subtitles data that are slightly positive.

```
NYT_sentiments_pos_neg_lower <- Normal_title_NYT_clean |>
  inner_join(get_sentiments("bing"))
```

```
Joining with `by = join_by(word)`
```

```
NYT_sentiments_pos_neg_lower
```

```
# A tibble: 1,242 x 6
   Article_ID Date       Subject                       Topic.Code word  sentiment
        <int> <date>     <fct>                              <int> <chr> <chr>
 1      41246 1996-01-01 Jails overwhelmed with hard~          12 stru~ negative
 2      41268 1996-01-03 Contenders for 1996 Presede~          20 cost~ negative
 3      41268 1996-01-03 Contenders for 1996 Presede~          20 plot  negative
 4      41279 1996-01-04 Bosnian Serb leader critici~          19 top   positive
```

```
 5      41279 1996-01-04 Bosnian Serb leader critici~         19 atta~ negative
 6      41302 1996-01-07 political violence in south~         19 stum~ negative
 7      41302 1996-01-07 political violence in south~         19 riva~ negative
 8      41314 1996-01-08 economists not afraid of a ~          1 fear  negative
 9      41344 1996-01-11 census changes                       20 cool  positive
10      41418 1996-01-19 uneasy time in saudi arabia          19 unea~ negative
# i 1,232 more rows
```

```
NYT_sentiments_pos_neg_upper <- UppercaseNYT_clean |>
  inner_join(get_sentiments("bing"))
```

```
Joining with `by = join_by(word)`
```

```
NYT_sentiments_pos_neg_upper
```

```
# A tibble: 1,050 x 6
   Article_ID Date       Subject                      Topic.Code word  sentiment
        <int> <date>     <fct>                             <int> <chr> <chr>
 1      41257 1996-01-02 Federal budget impasse affe~         20 impa~ negative
 2      41257 1996-01-02 Federal budget impasse affe~         20 inde~ negative
 3      41355 1996-01-12 barneys seeks bankruptcy             15 turm~ negative
 4      41355 1996-01-12 barneys seeks bankruptcy             15 diff~ negative
 5      41367 1996-01-14 clinton in bosnia                    19 thank positive
 6      41379 1996-01-15 AIDS HMOs                             3 trou~ negative
 7      41428 1996-01-20 Democrat Representative dis~         20 lure  negative
 8      41518 1996-01-30 new york governor and crime          24 crime negative
 9      41529 1996-01-31 lots of immigrants try to g~          5 futi~ negative
10      41529 1996-01-31 lots of immigrants try to g~          5 desp~ negative
# i 1,040 more rows
```

```
NYT_sentiments_pos_neg_post_colon <- PostColon_clean |>
  inner_join(get_sentiments("bing"))
```

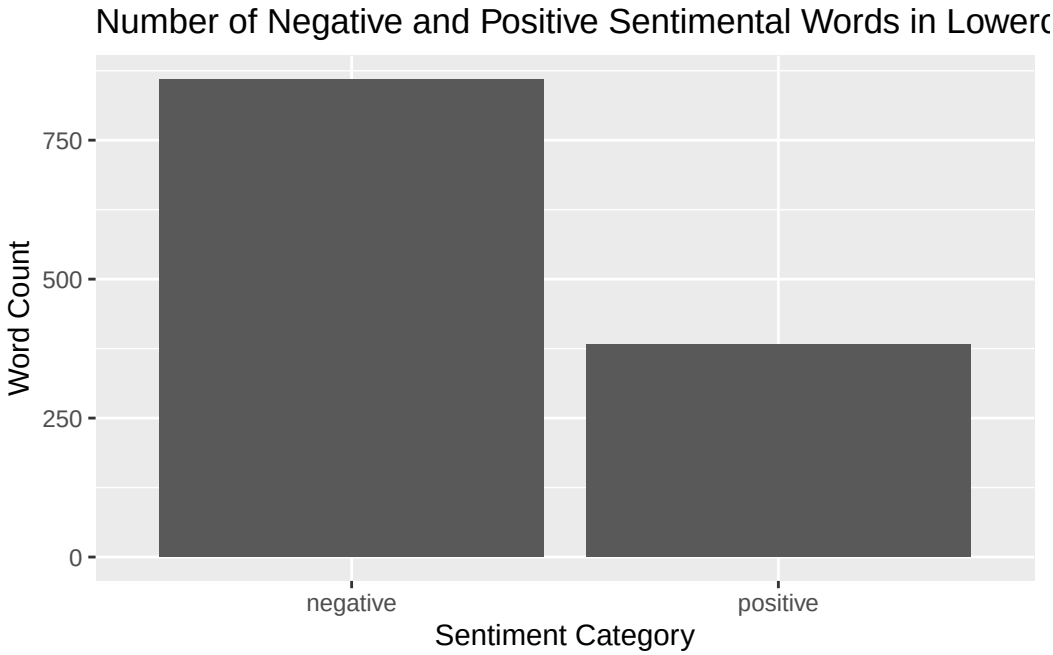
```
Joining with `by = join_by(word)`
```

```
NYT_sentiments_pos_neg_post_colon
```

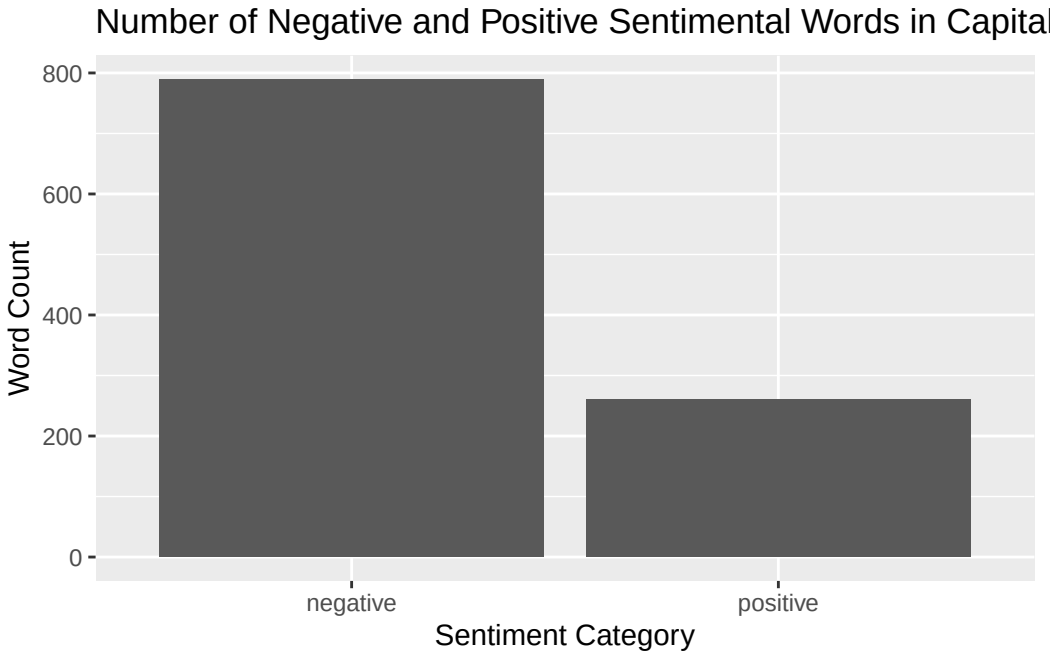
```
# A tibble: 132 x 6
   Article_ID Date       Subject                          Topic.Code word  sentiment
        <int> <date>     <fct>                                 <int> <chr> <chr>
 1      41355 1996-01-12 barneys seeks bankruptcy                 15 diff~ negative
 2      41428 1996-01-20 Democrat Representative dis~             20 lure  negative
 3      41621 1996-02-11 marriage in japan                        19 bett~ positive
 4      41621 1996-02-11 marriage in japan                        19 worse negative
 5      41621 1996-02-11 marriage in japan                        19 love  positive
 6      41792 1996-02-29 south carolina primary                   20 issu~ negative
 7      41815 1996-03-03 dole wins in south carolina              20 win   positive
 8      41883 1996-03-11 pat buchanan campaign                    20 issu~ negative
 9      41883 1996-03-11 pat buchanan campaign                    20 fun   positive
10      42085 1996-04-01 dick armey                               20 gruff negative
# i 122 more rows
```

Now let's move onto sentimental data where instead of assigning positive and negative sentimental values, we only look at positive and negative sentimental categories.

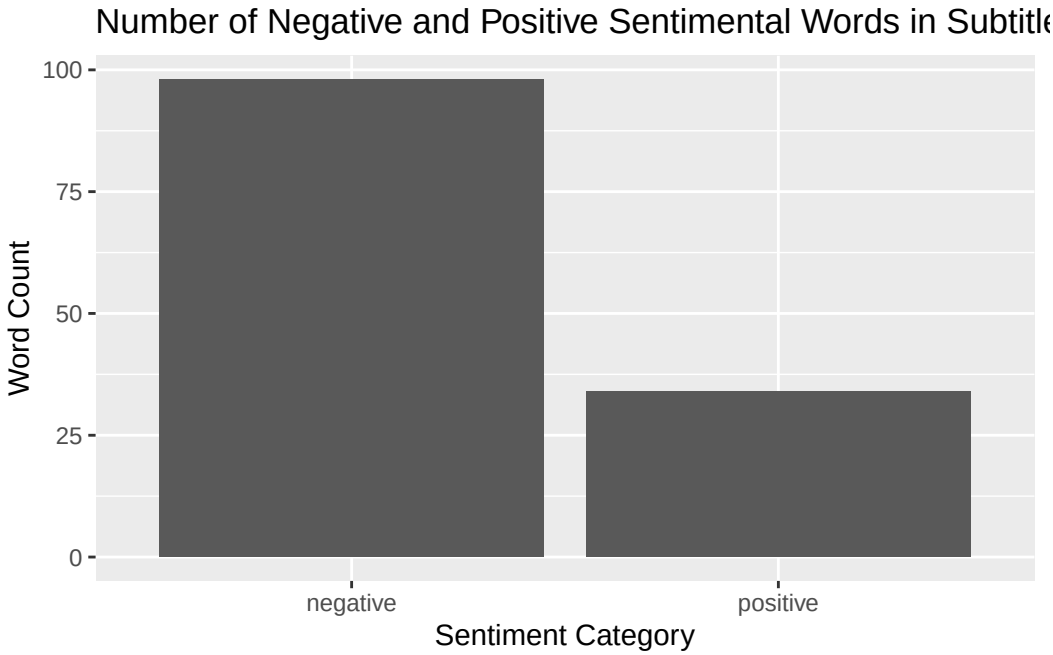
```
NYT_sentiments_pos_neg_lower |>
  mutate(year = str_sub(Date, 1, 4)) |>
  ggplot(aes(x = sentiment)) +
  geom_bar() +
  labs(title = "Number of Negative and Positive Sentimental Words in Lowercase Titles", x
```




```
NYT_sentiments_pos_neg_upper |>
  mutate(year = str_sub(Date, 1, 4)) |>
  ggplot(aes(x = sentiment)) +
  geom_bar() +
  labs(title = "Number of Negative and Positive Sentimental Words in Capitalized Titles",
```



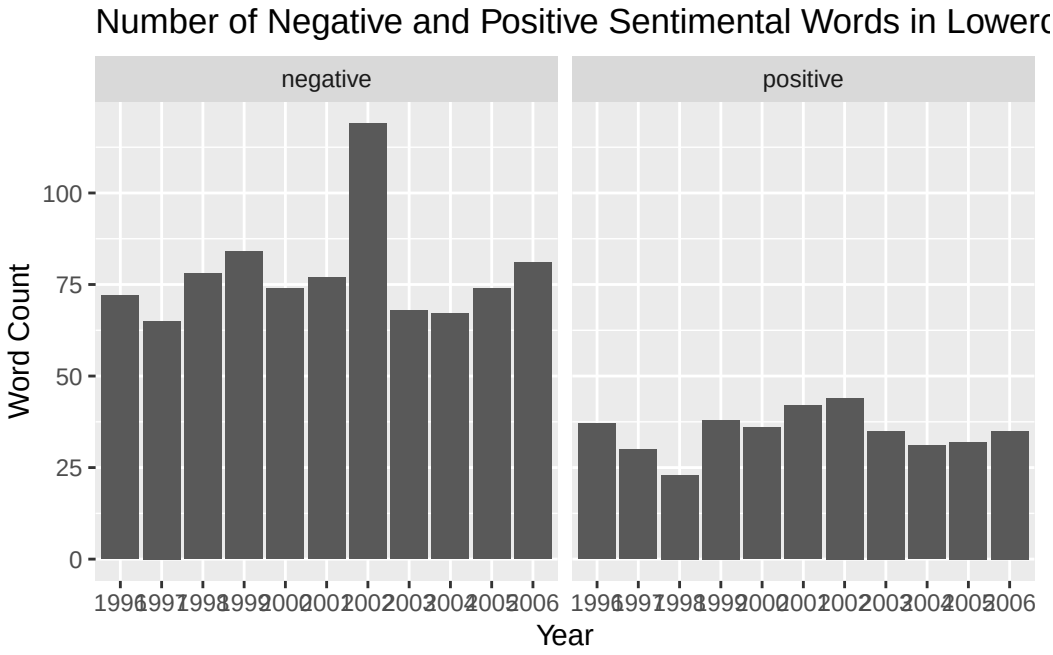

```
NYT_sentiments_pos_neg_post_colon |>
  mutate(year = str_sub(Date, 1, 4)) |>
  ggplot(aes(x = sentiment)) +
  geom_bar() +
  labs(title = "Number of Negative and Positive Sentimental Words in Subtitles", x = "Sent
```



The story that these plots tell showing the number of negative and positive sentimental words in each of the categories that we see is that speaking proportionally, there are far more negative, sentimental words being used in the titles containing uppercase words.

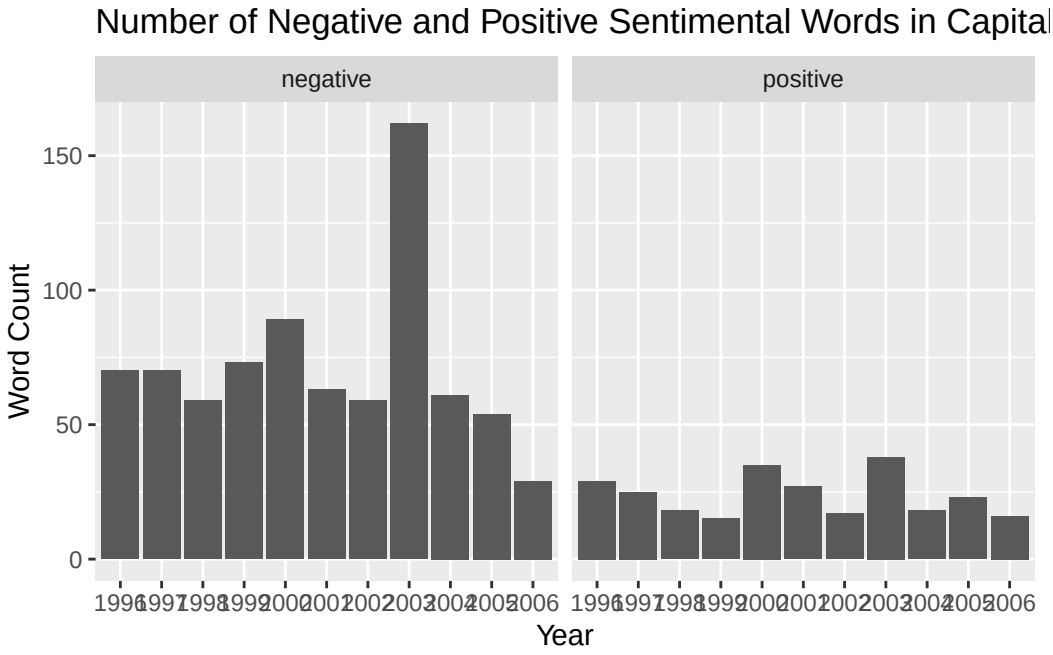
```
NYT_sentiments_pos_neg_lower |>
  mutate(year = str_sub(Date, 1, 4)) |>
  group_by(sentiment, year) |>
  summarize(n =n()) |>
  ungroup() |>
  ggplot(aes(x=year, y = n)) +
  geom_col() +
  facet_wrap(~sentiment) +
  labs(title = "Number of Negative and Positive Sentimental Words in Lowercase Title by Ye
```

```
`summarise()` has grouped output by 'sentiment'. You can override using the
`.groups` argument.
```



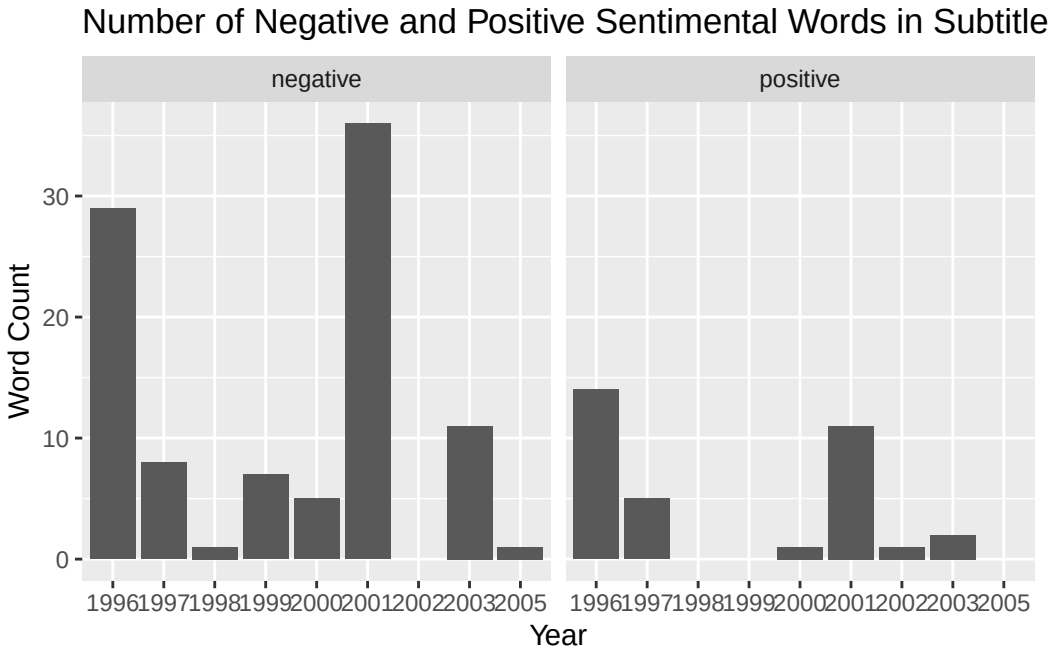

```
NYT_sentiments_pos_neg_upper |>
  mutate(year = str_sub(Date, 1, 4)) |>
  group_by(sentiment, year) |>
  summarize(n =n()) |>
  ungroup() |>
  ggplot(aes(x=year, y = n)) +
  geom_col() +
  facet_wrap(~sentiment) +
  labs(title = "Number of Negative and Positive Sentimental Words in Capitalized Title by
```

```
`summarise()` has grouped output by 'sentiment'. You can override using the
`.groups` argument.
```




```
NYT_sentiments_pos_neg_post_colon |>
  mutate(year = str_sub(Date, 1, 4)) |>
  group_by(sentiment, year) |>
  summarize(n =n()) |>
  ungroup() |>
  ggplot(aes(x=year, y = n)) +
  geom_col() +
  facet_wrap(~sentiment) +
  labs(title = "Number of Negative and Positive Sentimental Words in Subtitle by Years", x
```

```
`summarise()` has grouped output by 'sentiment'. You can override using the
`.groups` argument.
```



When looking at the same situation by year, it seems that the lowercase words were a preference for sentimental titles in 2002 while the uppercase words seem to be a preference for sentimental titles in 2003. The subtitles apparently tended to be more sentimental in the years 1996 and 2001. This is true for all the years stated for all these categories for both the positive and negative sentiments.

```
NYT_sentiments_lower <- Normal_title_NYT_clean |>
  inner_join(get_sentiments("nrc"))
```

```
Joining with `by = join_by(word)`
```

```
Warning in inner_join(Normal_title_NYT_clean, get_sentiments("nrc")): Detected an unexpected
i Row 4 of `x` matches multiple rows in `y`.
i Row 2672 of `y` matches multiple rows in `x`.
i If a many-to-many relationship is expected, set `relationship =
  "many-to-many"` to silence this warning.
```

```
NYT_sentiments_lower
```

```
# A tibble: 5,504 x 6
```

```
   Article_ID Date       Subject                       Topic.Code word  sentiment
        <int> <date>     <fct>                              <int> <chr> <chr>
 1      41246 1996-01-01 Jails overwhelmed with hard~          12 stru~ anger
 2      41246 1996-01-01 Jails overwhelmed with hard~          12 stru~ fear
 3      41246 1996-01-01 Jails overwhelmed with hard~          12 stru~ negative
 4      41246 1996-01-01 Jails overwhelmed with hard~          12 stru~ sadness
 5      41246 1996-01-01 Jails overwhelmed with hard~          12 surge surprise
 6      41268 1996-01-03 Contenders for 1996 Presede~          20 long  anticipa~
 7      41279 1996-01-04 Bosnian Serb leader critici~          19 top   anticipa~
 8      41279 1996-01-04 Bosnian Serb leader critici~          19 top   positive
 9      41279 1996-01-04 Bosnian Serb leader critici~          19 top   trust
10      41279 1996-01-04 Bosnian Serb leader critici~          19 lead~ positive
# i 5,494 more rows
```

```
NYT_sentiments_upper <- UppercaseNYT_clean |>
  inner_join(get_sentiments("nrc"))
```

```
Joining with `by = join_by(word)`

Warning in inner_join(UppercaseNYT_clean, get_sentiments("nrc")): Detected an unexpected many
i Row 7 of `x` matches multiple rows in `y`.
i Row 1118 of `y` matches multiple rows in `x`.
i If a many-to-many relationship is expected, set `relationship =
  "many-to-many"` to silence this warning.
```

```
NYT_sentiments_upper
```

```
# A tibble: 5,081 x 6
   Article_ID Date       Subject                       Topic.Code word  sentiment
        <int> <date>     <fct>                              <int> <chr> <chr>
 1      41257 1996-01-02 Federal budget impasse affe~          20 inde~ negative
 2      41290 1996-01-05 Battle over budget: Republi~          1 batt~ anger
 3      41290 1996-01-05 Battle over budget: Republi~          1 batt~ negative
 4      41290 1996-01-05 Battle over budget: Republi~          1 budg~ trust
 5      41290 1996-01-05 Battle over budget: Republi~          1 plan  anticipa~
 6      41333 1996-01-10 budget fight                          1 batt~ anger
 7      41333 1996-01-10 budget fight                          1 batt~ negative
 8      41333 1996-01-10 budget fight                          1 budg~ trust
 9      41333 1996-01-10 budget fight                          1 budg~ trust
10      41355 1996-01-12 barneys seeks bankruptcy             15 turm~ anger
# i 5,071 more rows
```

```
NYT_sentiments_post_colon <- PostColon_clean |>
  inner_join(get_sentiments("nrc"))
```

```
Joining with `by = join_by(word)`
```

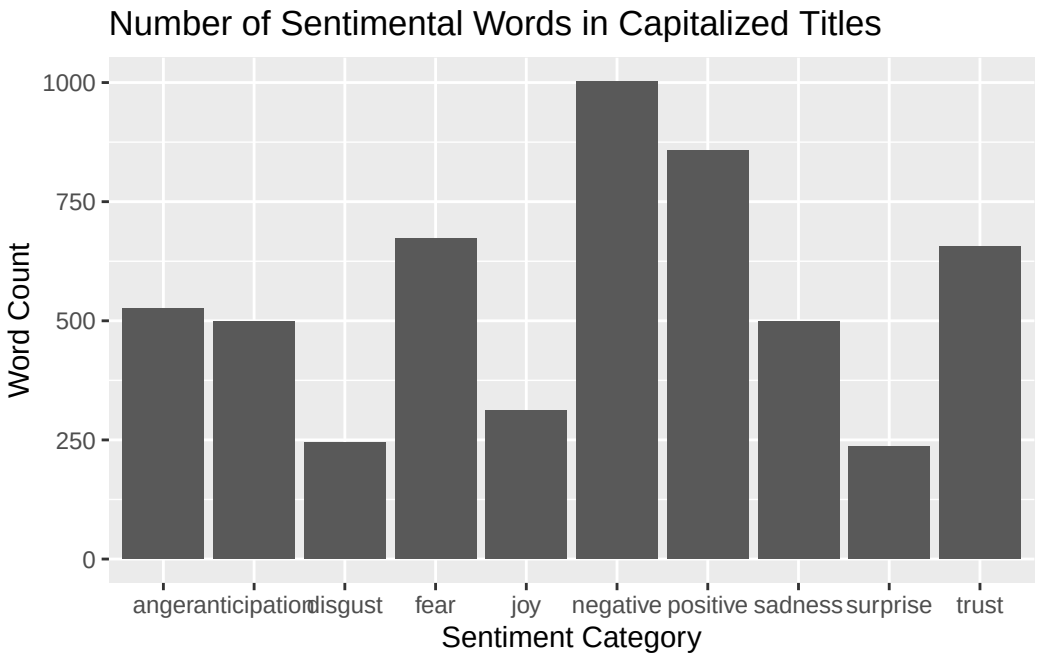
```
Warning in inner_join(PostColon_clean, get_sentiments("nrc")): Detected an unexpected many-to
i Row 24 of `x` matches multiple rows in `y`.
i Row 4029 of `y` matches multiple rows in `x`.
i If a many-to-many relationship is expected, set `relationship =
  "many-to-many"` to silence this warning.
```

```
NYT_sentiments_post_colon
```

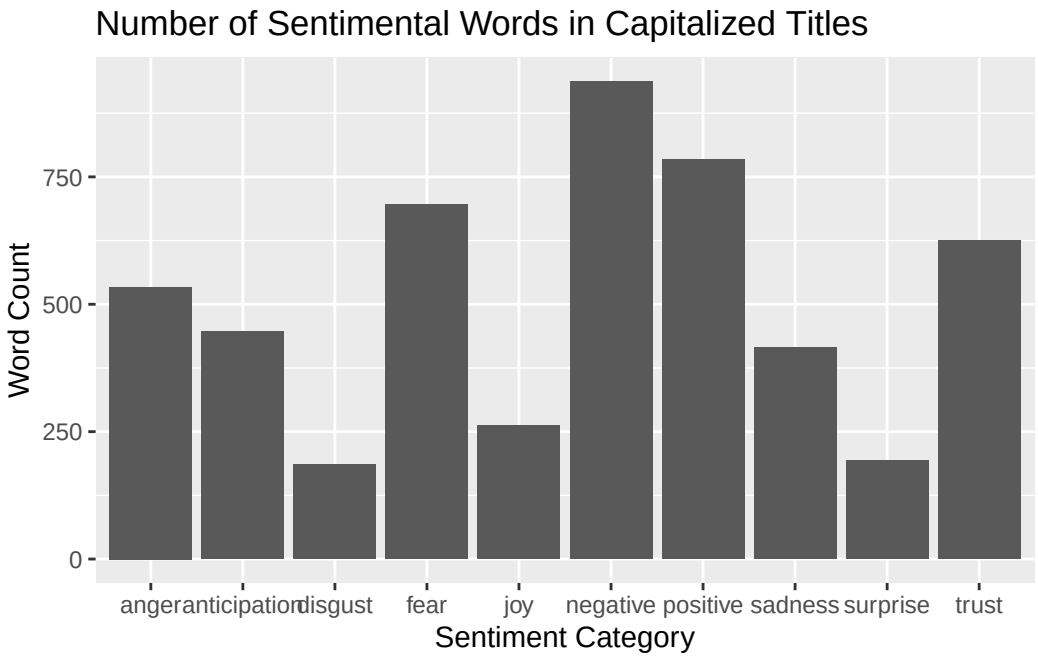
```
# A tibble: 877 x 6
   Article_ID Date       Subject                       Topic.Code word  sentiment
        <int> <date>     <fct>                              <int> <chr> <chr>
 1      41290 1996-01-05 Battle over budget: Republi~           1 plan  anticipa~
 2      41333 1996-01-10 budget fight                            1 budg~ trust
 3      41355 1996-01-12 barneys seeks bankruptcy               15 diff~ negative
 4      41355 1996-01-12 barneys seeks bankruptcy               15 diff~ sadness
 5      41355 1996-01-12 barneys seeks bankruptcy               15 bank~ anger
 6      41355 1996-01-12 barneys seeks bankruptcy               15 bank~ disgust
 7      41355 1996-01-12 barneys seeks bankruptcy               15 bank~ fear
 8      41355 1996-01-12 barneys seeks bankruptcy               15 bank~ negative
 9      41355 1996-01-12 barneys seeks bankruptcy               15 bank~ sadness
10      41355 1996-01-12 barneys seeks bankruptcy               15 fight anger
# i 867 more rows
```

Now, let's move onto data containing sentiment categories.

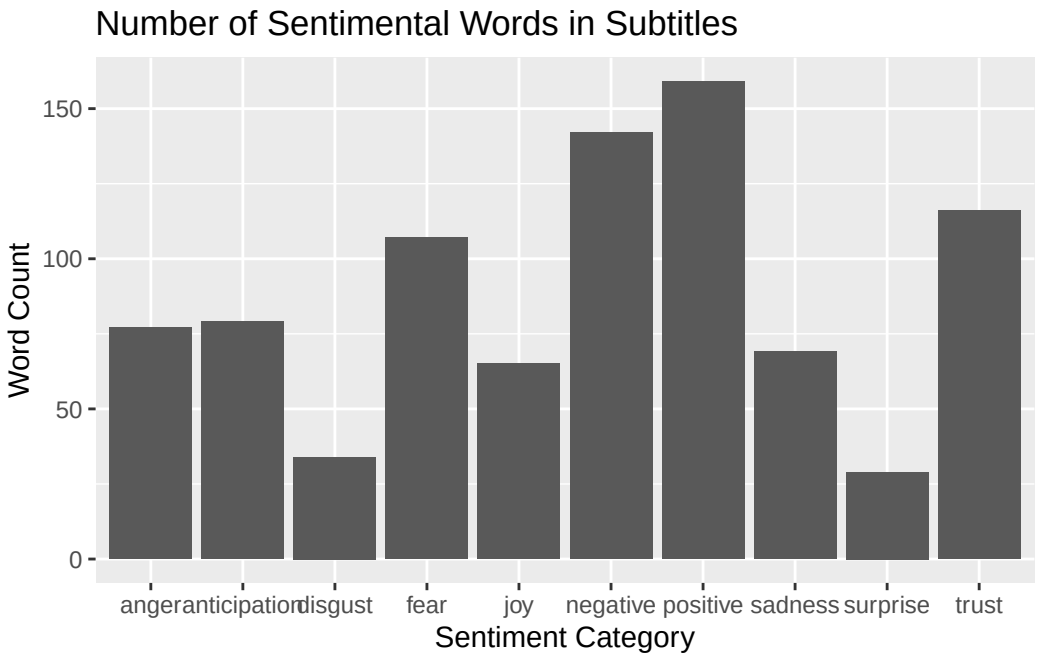
```
NYT_sentiments_lower |>
  mutate(year = str_sub(Date, 1, 4)) |>
  ggplot(aes(x = sentiment)) +
  geom_bar() +
  labs(title = "Number of Sentimental Words in Capitalized Titles", x = "Sentiment Categor
```




```
NYT_sentiments_upper |>
  mutate(year = str_sub(Date, 1, 4)) |>
  ggplot(aes(x = sentiment)) +
  geom_bar() +
  labs(title = "Number of Sentimental Words in Capitalized Titles", x = "Sentiment Categor
```



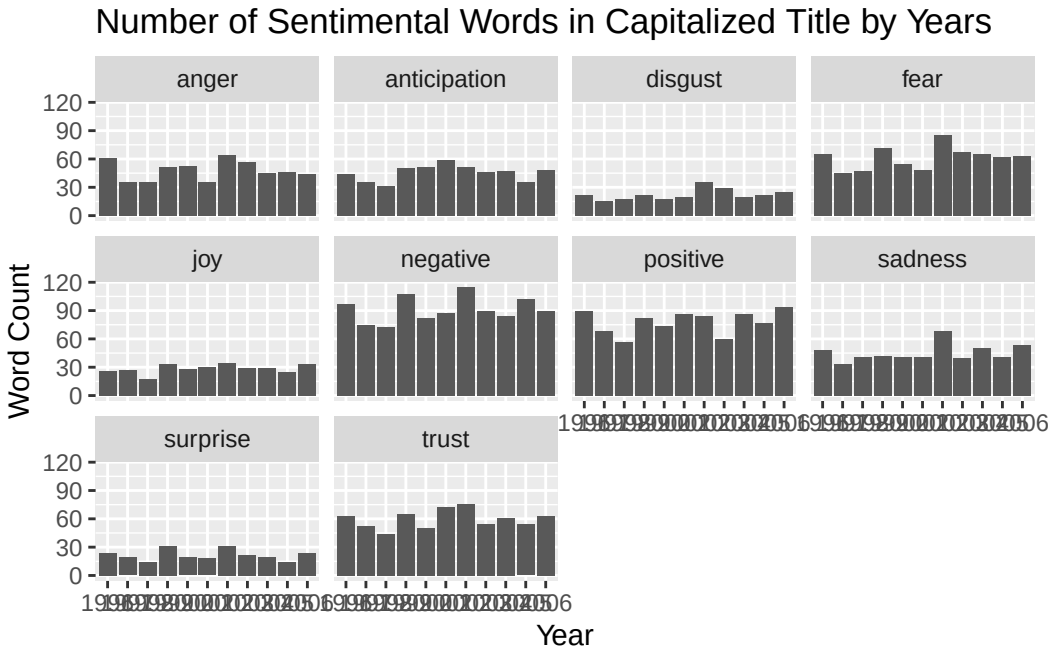

```
NYT_sentiments_post_colon |>
  mutate(year = str_sub(Date, 1, 4)) |>
  ggplot(aes(x = sentiment)) +
  geom_bar() +
  labs(title = "Number of Sentimental Words in Subtitles", x = "Sentiment Category", y = "
```



Fear sadness anger and disgust highest on same years. Little Joy and surprise.

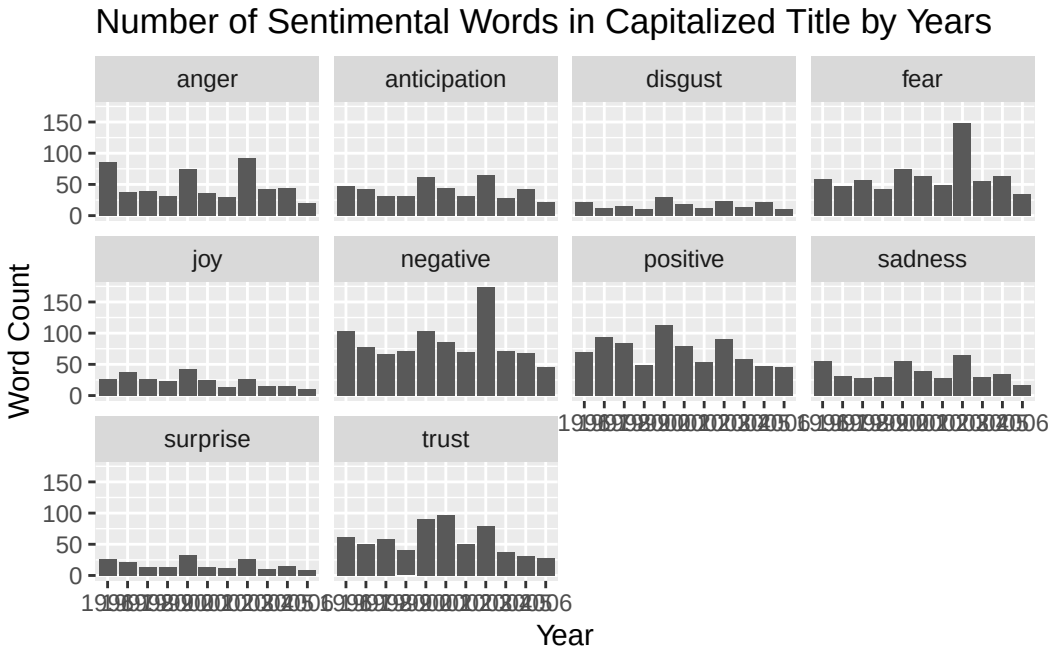
```
NYT_sentiments_lower |>
  mutate(year = str_sub(Date, 1, 4)) |>
  group_by(sentiment, year) |>
  summarize(n =n()) |>
  ungroup() |>
  ggplot(aes(x=year, y = n)) +
  geom_col() +
  facet_wrap(~sentiment) +
  labs(title = "Number of Sentimental Words in Capitalized Title by Years", x = "Year", y
```

```
`summarise()` has grouped output by 'sentiment'. You can override using the
`.groups` argument.
```



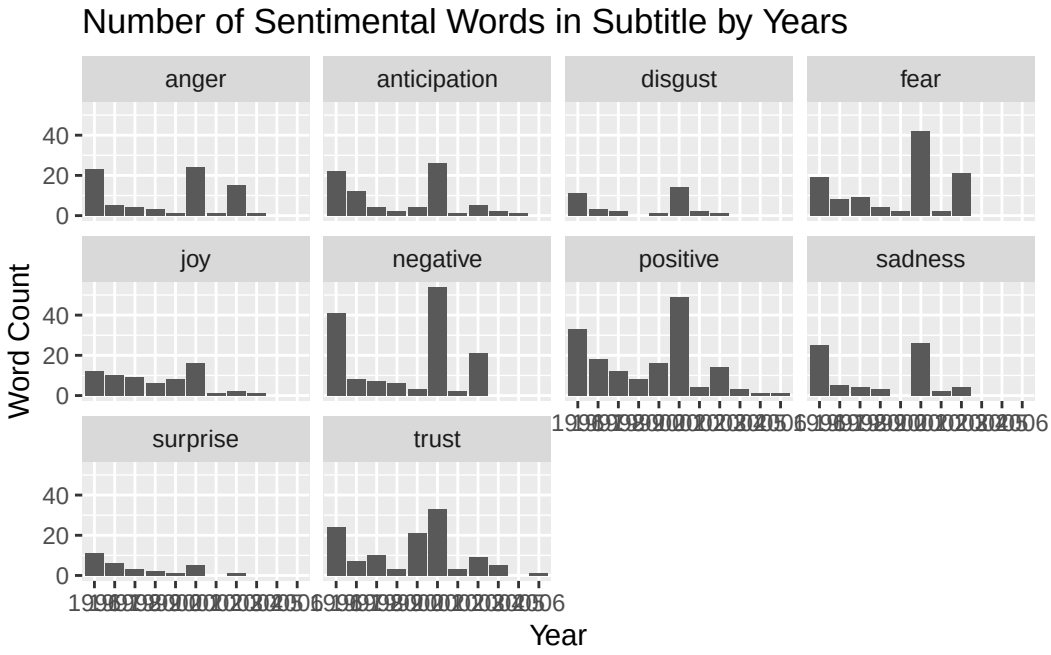

```
NYT_sentiments_upper |>
  mutate(year = str_sub(Date, 1, 4)) |>
  group_by(sentiment, year) |>
  summarize(n =n()) |>
  ungroup() |>
  ggplot(aes(x=year, y = n)) +
  geom_col() +
  facet_wrap(~sentiment) +
  labs(title = "Number of Sentimental Words in Capitalized Title by Years", x = "Year", y
```

```
`summarise()` has grouped output by 'sentiment'. You can override using the
`.groups` argument.
```




```
NYT_sentiments_post_colon |>
  mutate(year = str_sub(Date, 1, 4)) |>
  group_by(sentiment, year) |>
  summarize(n =n()) |>
  ungroup() |>
  ggplot(aes(x=year, y = n)) +
  geom_col() +
  facet_wrap(~sentiment) +
  labs(title = "Number of Sentimental Words in Subtitle by Years", x = "Year", y = "Word C
```

```
`summarise()` has grouped output by 'sentiment'. You can override using the
`.groups` argument.
```



For both the uppercase and lowercase data it seems like the sentiments categories of fear sadness anger and discussed are all very high on the same years. It is also worth noting that there is very little joy and surprise in all these years further showing the news does tend to be negative in terms of sentiments as predicted. It is tough to analyze this data for the subtitles since we have a small sample size and most of these sentiment categories seem to look similar. However, we still have the same trends of little joy and surprise as in the other two examples.

Conclusion

Uppercase titles tend to be more negative than lowercase titles while even though the subtitles with uppercase words tend to be more neutral than the other two categories, they are still slightly negative.