Sentiment Analysis of COVID-19 Tweets

2024-11-11

Introduction: The purpose of this assignment is to conduct sentiment analysis. I decided to analyze a collection of tweets to explore and quantify the emotional tone of social media interactions during COVID-19. Sentiment analysis is a natural language processing technique used to classify text by identifying and categorizing emotions or attitudes conveyed within the language. I will beapplying different sentiment lexicons — namely, AFINN, Bing, and NRC.

The AFINN lexicon assigns sentiment scores, allowing a calculation of cumulative sentiment per tweet, while the Bing lexicon categorizes words simply as "positive" or "negative." The NRC lexicon provides a broader emotional classification, encompassing various sentiment categories.

```r
library(tidytext)

get_sentiments("afinn")
```

```
## # A tibble: 2,477 x 2
##    word       value
##    <chr>      <dbl>
##  1 abandon       -2
##  2 abandoned     -2
##  3 abandons      -2
##  4 abducted      -2
##  5 abduction     -2
##  6 abductions    -2
##  7 abhor         -3
##  8 abhorred      -3
##  9 abhorrent     -3
## 10 abhors        -3
## # i 2,467 more rows
```

```r
get_sentiments("bing")
```

```
## # A tibble: 6,786 x 2
##    word        sentiment
##    <chr>       <chr>
##  1 2-faces     negative
##  2 abnormal    negative
##  3 abolish     negative
##  4 abominable  negative
##  5 abominably  negative
##  6 abominate   negative
##  7 abomination negative
##  8 abort       negative
##  9 aborted     negative
## 10 aborts      negative
## # i 6,776 more rows
```

```r
get_sentiments("nrc")
```

```
## # A tibble: 13,872 x 2
##    word        sentiment
##    <chr>       <chr>
##  1 abacus      trust
##  2 abandon     fear
##  3 abandon     negative
##  4 abandon     sadness
##  5 abandoned   anger
##  6 abandoned   fear
##  7 abandoned   negative
##  8 abandoned   sadness
##  9 abandonment anger
## 10 abandonment fear
## # i 13,862 more rows
```

```r
# Load Libraries
library(rtweet)
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
```

```r
library(ggplot2)
```

```r
# Load the dataset
tweets <- read.csv("/Users/leslietavarez/Downloads/covid19_tweets.csv")

# Preview the data
head(tweets)
```

```
##           user_name        user_location
## 1                         astroworld
## 2     Tom Basile           New York, NY
## 3   Time4fisticuffs     Pewee Valley, KY
## 4       ethel mertz Stuck in the Middle
## 5          DIPR-J&K    Jammu and Kashmir
## 6   Franz Schubert              ´
##
## 1                                                                                wednesday addams as
## 2 Husband, Father, Columnist & Commentator. Author of Tough Sell: Fighting the Media War in Iraq. Bu
## 3                             #Christian #Catholic #Conservative #Reagan #Republican #Capitalist; Spe
## 4
## 5                                                        Official Twitter handle of Department of Inf
## 6                                                                        #   ´   #Novorossiya #
##          user_created user_followers user_friends user_favourites user_verified
## 1 2017-05-26 05:46:42            624          950           18775         False
## 2 2009-04-16 20:06:23           2253         1677              24          True
## 3 2009-02-28 18:57:41           9275         9525            7254         False
## 4 2019-03-07 01:45:06            197          987            1488         False
## 5 2017-02-12 06:45:15         101009          168             101         False
## 6 2018-03-19 16:29:52           1180         1071            1287         False
##                   date
## 1 2020-07-25 12:27:21
## 2 2020-07-25 12:27:17
## 3 2020-07-25 12:27:14
## 4 2020-07-25 12:27:10
## 5 2020-07-25 12:27:08
## 6 2020-07-25 12:27:06
##
## 1 If I smelled the scent of hand sanitizers today on someone in the past, I would think they were so
## 2 Hey @Yankees @YankeesPR and @MLB - wouldn't it have made more sense to have the players pay their
## 3 @diane3443 @wdunlap @realDonaldTrump Trump never once claimed #COVID19 was a hoax. We all claim tha
## 4  @brookbanktv The one gift #COVID19 has give me is an appreciation for the simple things that were
## 5  25 July : Media Bulletin on Novel #CoronaVirusUpdates #COVID19 \n@kansalrohit69 @DrSyedSehrish @a
## 6 #coronavirus #covid19 deaths continue to rise. It's almost  as bad as it ever was.  Politicians an
##                             hashtags              source is_retweet
## 1                                     Twitter for iPhone      False
## 2                                    Twitter for Android      False
## 3                       ['COVID19'] Twitter for Android      False
## 4                       ['COVID19']  Twitter for iPhone      False
## 5 ['CoronaVirusUpdates', 'COVID19'] Twitter for Android      False
## 6        ['coronavirus', 'covid19']     Twitter Web App      False
```

```
# Check structure and add unique IDs to each tweet row
tweets_clean <- tweets %>%
  mutate(tweet_id = row_number()) %>% # Creates a unique ID for each tweet row
 select(tweet_id, text)

# Tokenize text by individual tweet and remove stop words
tweets_words <- tweets_clean %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words, by = "word")

head(tweets_clean)
```

```
##   tweet_id
## 1        1
## 2        2
## 3        3
## 4        4
## 5        5
## 6        6
##
## 1 If I smelled the scent of hand sanitizers today on someone in the past, I would think they were so
## 2 Hey @Yankees @YankeesPR and @MLB - wouldn't it have made more sense to have the players pay their
## 3 @diane3443 @wdunlap @realDonaldTrump Trump never once claimed #COVID19 was a hoax. We all claim tha
## 4  @brookbanktv The one gift #COVID19 has give me is an appreciation for the simple things that were
## 5  25 July : Media Bulletin on Novel #CoronaVirusUpdates #COVID19 \n@kansalrohit69 @DrSyedSehrish @a
## 6 #coronavirus #covid19 deaths continue to rise. It's almost  as bad as it ever was.  Politicians an
```

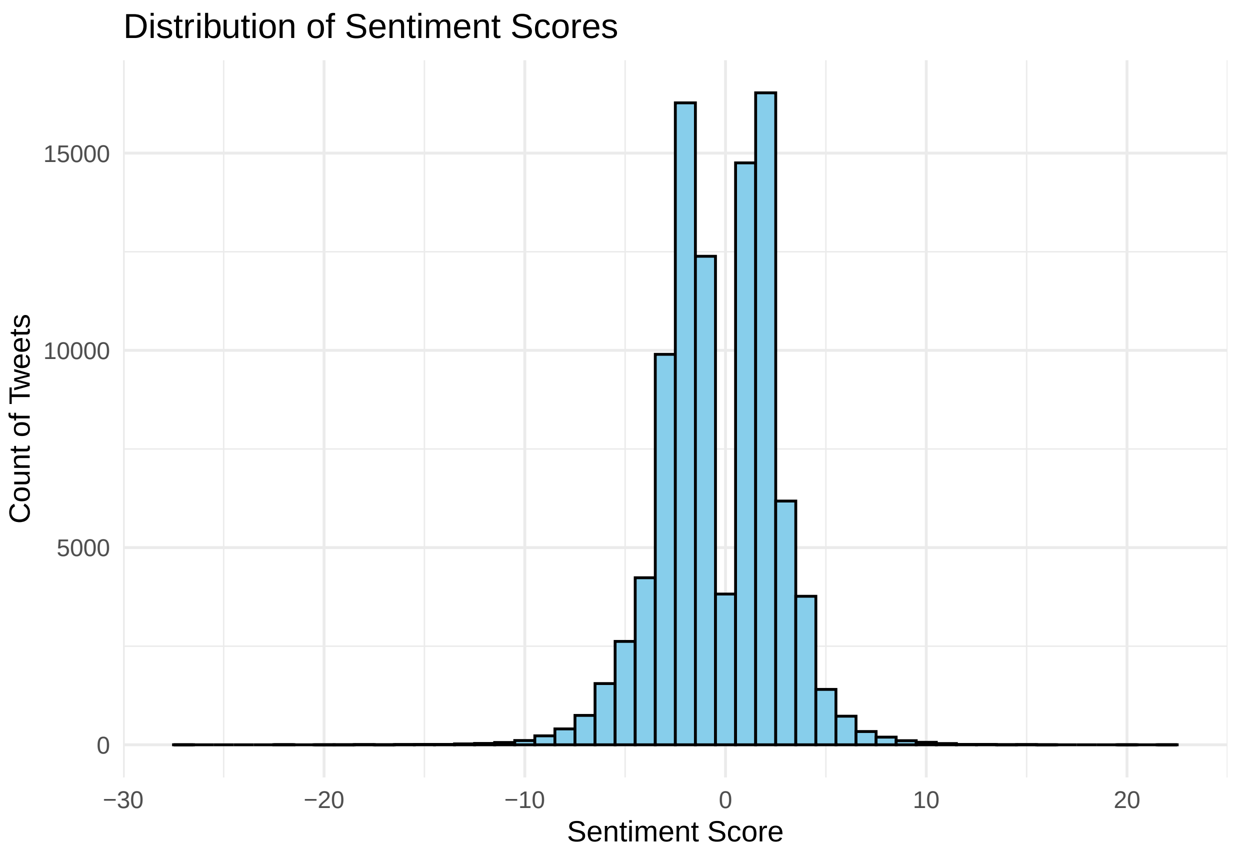
The histogram from the AFINN lexicon reveals a nearly symmetric distribution around zero, indicating a balanced mix of positive and negative sentiment in the COVID-19 tweets.

```r
# Perform sentiment analysis using AFINN lexicon
tweets_sentiment <- tweets_words %>%
  inner_join(get_sentiments("afinn"), by = "word") %>%
  group_by(tweet_id) %>%  # assuming 'tweet_id' is a unique identifier for each tweet
  summarize(sentiment_score = sum(value, na.rm = TRUE))

#View results
tweets_sentiment
```

```
## # A tibble: 96,523 x 2
##    tweet_id sentiment_score
##       <int>           <dbl>
##  1        2              -1
##  2        3              -2
##  3        4               4
##  4        6              -3
##  5        9               1
##  6       10               1
##  7       11              -1
##  8       13               4
##  9       14               1
## 10       19              -2
## # i 96,513 more rows
```

```r
# Histogram of sentiment scores
ggplot(tweets_sentiment, aes(x = sentiment_score)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Sentiment Scores",
       x = "Sentiment Score",
       y = "Count of Tweets") +
  theme_minimal()
```



The positive sentiment is surprisingly strong, with more positive than negative tweets overall. However, there's a noticeable mix of emotions—anticipation, fear, sadness, and trust all show up frequently. It's interesting to see so many positive tweets, as I expected more negativity considering how hard the pandemic was for many, with people losing loved ones and facing uncertainty. This mix of sentiments really captures the complex emotions people felt during that time.

```
# Perform sentiment analysis using NRC lexicon
tweets_sentiment_nrc <- tweets_words %>%
  inner_join(get_sentiments("nrc"), by = "word") %>%
  group_by(tweet_id, sentiment) %>%  # Group by tweet and sentiment type
  summarize(sentiment_count = n(), .groups = 'drop')  # Count occurrences of each sentiment per tweet
```

```
## Warning in inner_join(., get_sentiments("nrc"), by = "word"): Detected an unexpected many-to-many rel
## i Row 6 of `x` matches multiple rows in `y`.
## i Row 10794 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.
```

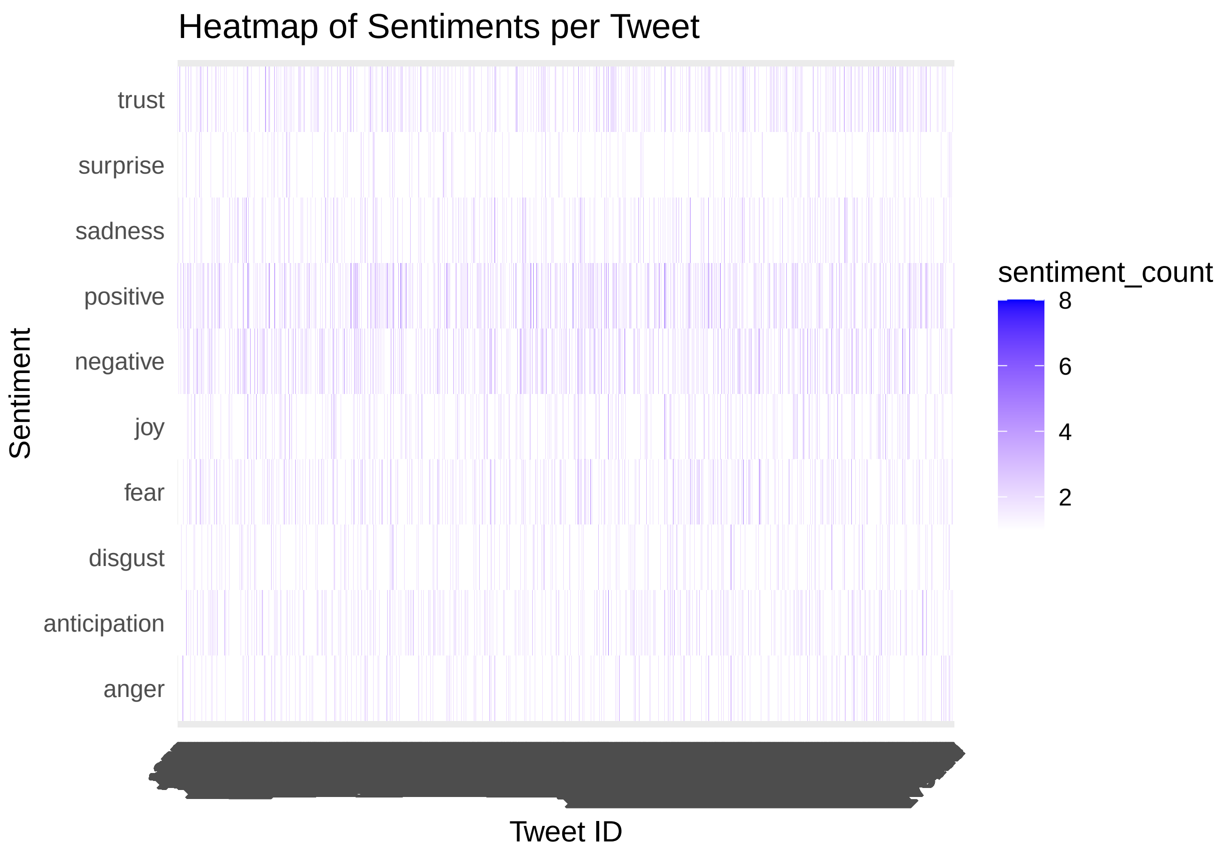
```
# View the results
head(tweets_sentiment_nrc)
```

```
## # A tibble: 6 x 3
##   tweet_id sentiment    sentiment_count
##      <int> <chr>                  <int>
## 1        1 disgust                    1
## 2        1 negative                   1
## 3        2 anticipation               1
## 4        2 joy                        1
## 5        2 positive                   3
## 6        2 trust                      2
```

```
tweets_sentiment_nrc
```

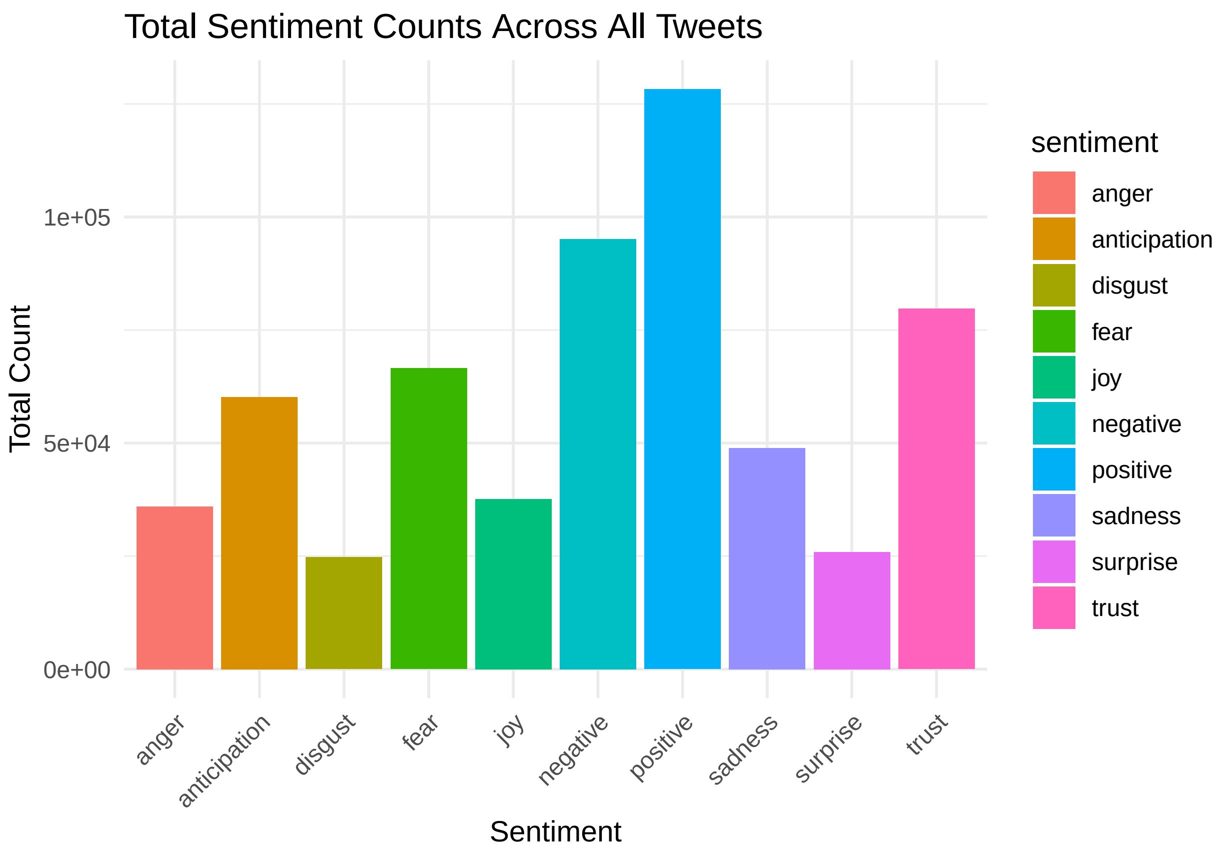
```
## # A tibble: 471,050 x 3
##    tweet_id sentiment    sentiment_count
##       <int> <chr>                  <int>
##  1        1 disgust                    1
##  2        1 negative                   1
##  3        2 anticipation               1
##  4        2 joy                        1
##  5        2 positive                   3
##  6        2 trust                      2
##  7        3 anger                      1
##  8        3 disgust                    1
##  9        3 negative                   1
## 10        3 positive                   1
## # i 471,040 more rows
```

```
# Heatmap of sentiment counts
ggplot(tweets_sentiment_nrc, aes(x = factor(tweet_id), y = sentiment, fill = sentiment_count)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "blue") +
  labs(title = "Heatmap of Sentiments per Tweet",
       x = "Tweet ID",
       y = "Sentiment") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```r
# Summarize total counts per sentiment
sentiment_summary <- tweets_sentiment_nrc %>%
  group_by(sentiment) %>%
  summarize(total_count = sum(sentiment_count))

# Bar plot
ggplot(sentiment_summary, aes(x = sentiment, y = total_count, fill = sentiment)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Sentiment Counts Across All Tweets",
       x = "Sentiment",
       y = "Total Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The Bing lexicon analysis reveals a predominance of negative sentiment over positive sentiment in the COVID-19 tweets.This reflects the challenging aspects of the pandemic.However, the presence of positive sentiment shows that some tweets conveyed hope or moments of appreciation.

```
# Perform sentiment analysis using the Bing lexicon
tweets_sentiment_bing <- tweets_words %>%
  inner_join(get_sentiments("bing"), by = "word") %>%  # Join with Bing lexicon
  group_by(tweet_id, sentiment) %>%                    # Group by tweet and sentiment type
  summarize(sentiment_count = n(), .groups = 'drop')   # Count positive/negative words for each tweet
```

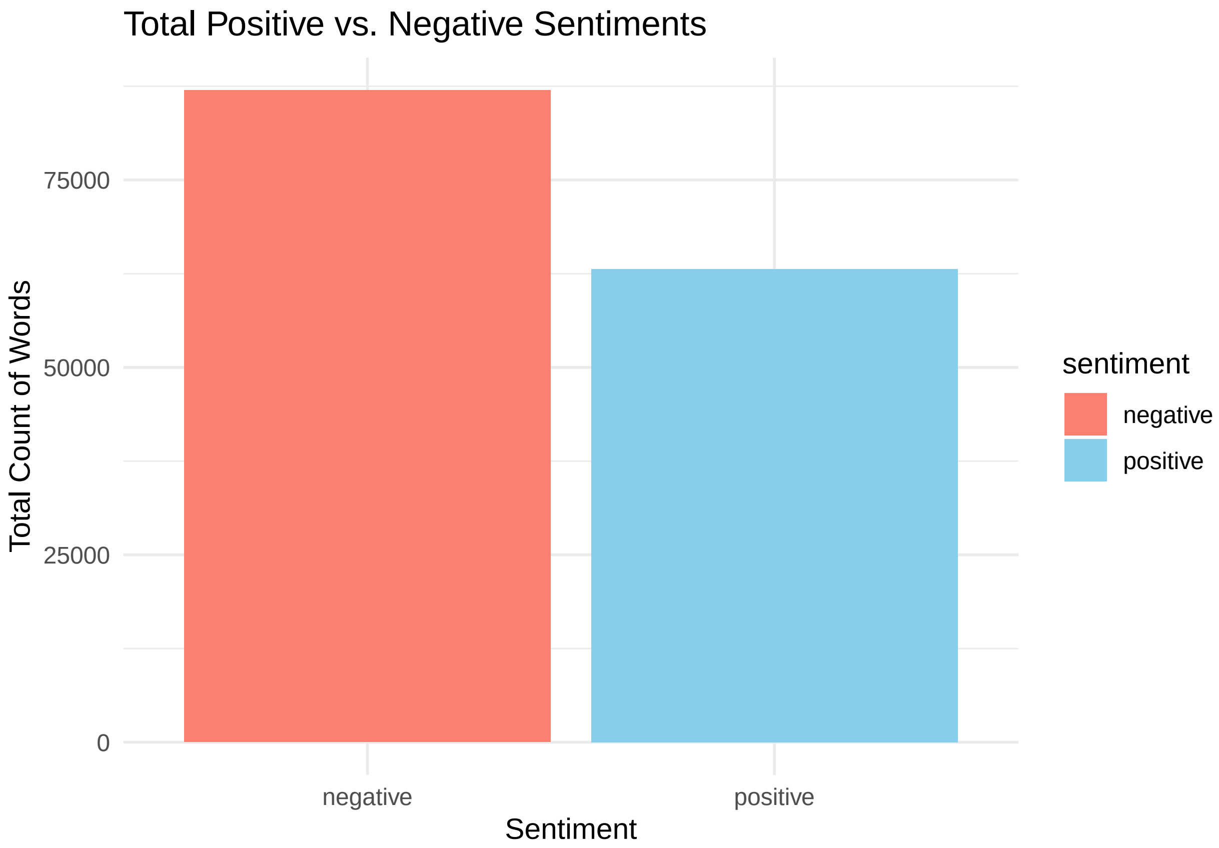
```
## Warning in inner_join(., get_sentiments("bing"), by = "word"): Detected an unexpected many-to-many r
## i Row 674776 of `x` matches multiple rows in `y`.
## i Row 5201 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```
# View the results
head(tweets_sentiment_bing)
```

```
## # A tibble: 6 x 3
##   tweet_id sentiment sentiment_count
##      <int> <chr>               <int>
## 1        1 negative                1
## 2        3 negative                1
## 3        3 positive                1
## 4        6 negative                  1
## 5        9 positive                  1
## 6       13 positive                  2
```

```r
# Summarize total positive and negative counts
sentiment_summary <- tweets_sentiment_bing %>%
  group_by(sentiment) %>%
  summarize(total_count = sum(sentiment_count))

# Bar plot for total positive vs. negative sentiment
ggplot(sentiment_summary, aes(x = sentiment, y = total_count, fill = sentiment)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = c("positive" = "skyblue", "negative" = "salmon")) +
  labs(title = "Total Positive vs. Negative Sentiments",
       x = "Sentiment",
       y = "Total Count of Words") +
  theme_minimal()
```



Conclusion:

The Bing lexicon appears more negative than others because it classifies words like “quarantine” or “isolation” as inherently negative, even though they might be neutral in context. Bing’s binary approach exaggerates negativity, especially in difficult topics like COVID-19. In contrast, NRC captures a broader range of emotions, leading to a more balanced distribution, while AFINN’s numerical scoring system allows for more nuance, offsetting mild negativity with extreme positivity for a more balanced sentiment analysis.