Sentiment Analysis with Tidy Data

Escarlet Gabriel Vicente

2025-11-02

1. Load Libraries

```
library(tidyverse)
library(tidytext)
library(janeaustenr)
library(gutenbergr)
library(lexicon)
library(ggplot2)
library(tidyr)
```

2. Jane Austen Example (Base)

In this section, I recreated the base example from Text Mining with R using Jane Austen's novels. I tokenized the text and applied sentiment analysis using the Bing and NRC lexicons.

```
tidy_books <- austen_books() %>%
group_by(book) %>%
mutate(
linenumber = row_number(),
chapter = cumsum(stringr::str_detect(
  text, stringr::regex("^chapter [\\divxlc]", ignore_case = TRUE)
))
) %>%
ungroup() %>%
unnest_tokens(word, text)
```

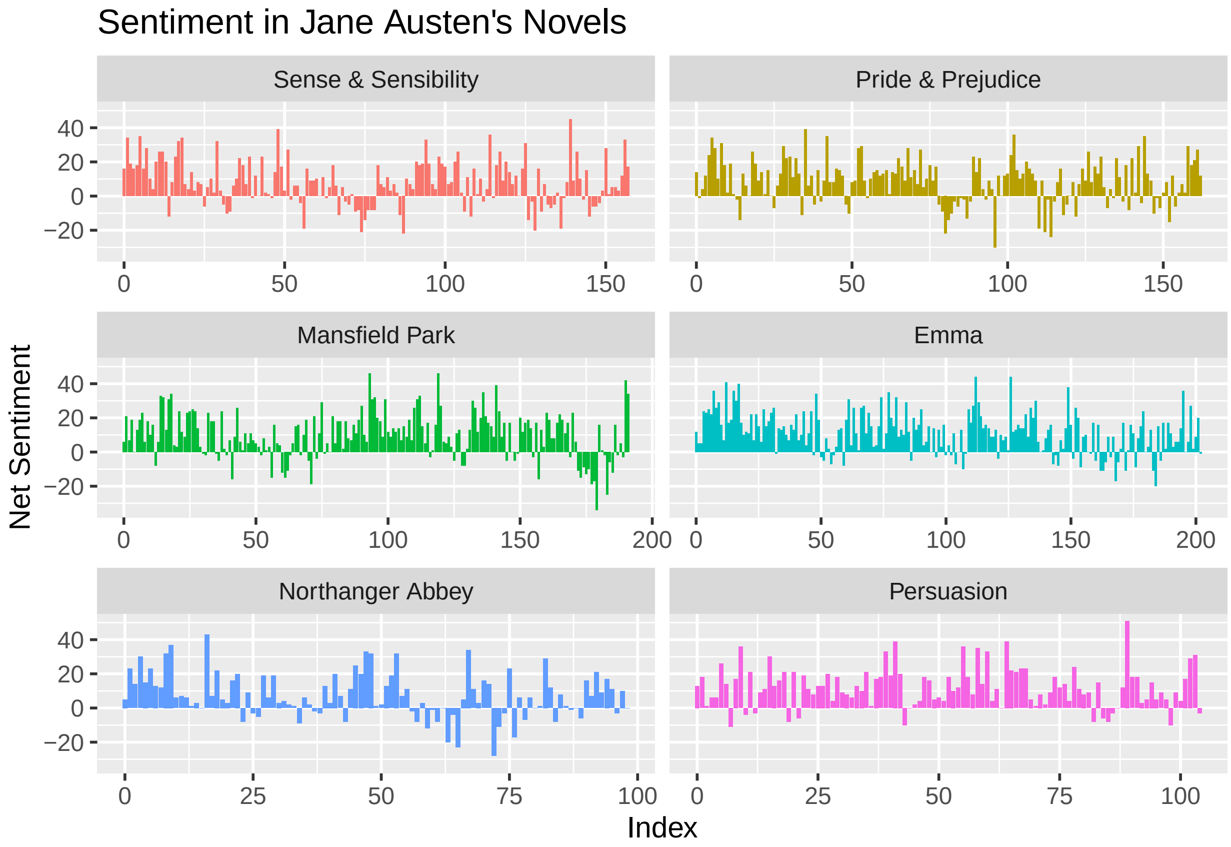
Top Joy Words in Emma (NRC Lexicon) Here, I explored which words were most commonly associated with joy in Emma according to the NRC lexicon.

```
nrc_joy <- get_sentiments("nrc") %>% filter(sentiment == "joy")

tidy_books %>%
filter(book == "Emma") %>%
inner_join(nrc_joy, by = "word") %>%
count(word, sort = TRUE) %>%
head(10)
```

```
## # A tibble: 10 x 2
##    word          n
##    <chr>     <int>
##  1 good        359
##  2 friend      166
##  3 hope        143
##  4 happy       125
##  5 love        117
##  6 deal         92
##  7 found        92
##  8 present      89
##  9 kind         82
## 10 happiness    76
```

Sentiment Over Narrative (Bing Lexicon) Next, I analyzed how the overall sentiment shifted throughout each novel using the Bing lexicon.

```
jane_austen_sentiment <- tidy_books %>%
inner_join(get_sentiments("bing"), by = "word") %>%
count(book, index = linenumber %/% 80, sentiment) %>%
pivot_wider(names_from = sentiment, values_from = n, values_fill = 0) %>%
mutate(sentiment = positive - negative)

ggplot(jane_austen_sentiment, aes(index, sentiment, fill = book)) +
geom_col(show.legend = FALSE) +
facet_wrap(~book, ncol = 2, scales = "free_x") +
labs(title = "Sentiment in Jane Austen's Novels",
x = "Index", y = "Net Sentiment")
```



3. Independent Analysis — Pride and Prejudice

I then applied the same text-mining workflow to a single novel, Pride and Prejudice, to perform my own sentiment analysis.

```
pride_text <- austen_books() %>%
  filter(book == "Pride & Prejudice") %>%
  unnest_tokens(word, text)

head(pride_text)
```

```
## # A tibble: 6 x 2
##   book              word
##   <fct>             <chr>
## 1 Pride & Prejudice pride
## 2 Pride & Prejudice and
## 3 Pride & Prejudice prejudice
## 4 Pride & Prejudice by
## 5 Pride & Prejudice jane
## 6 Pride & Prejudice austen
```

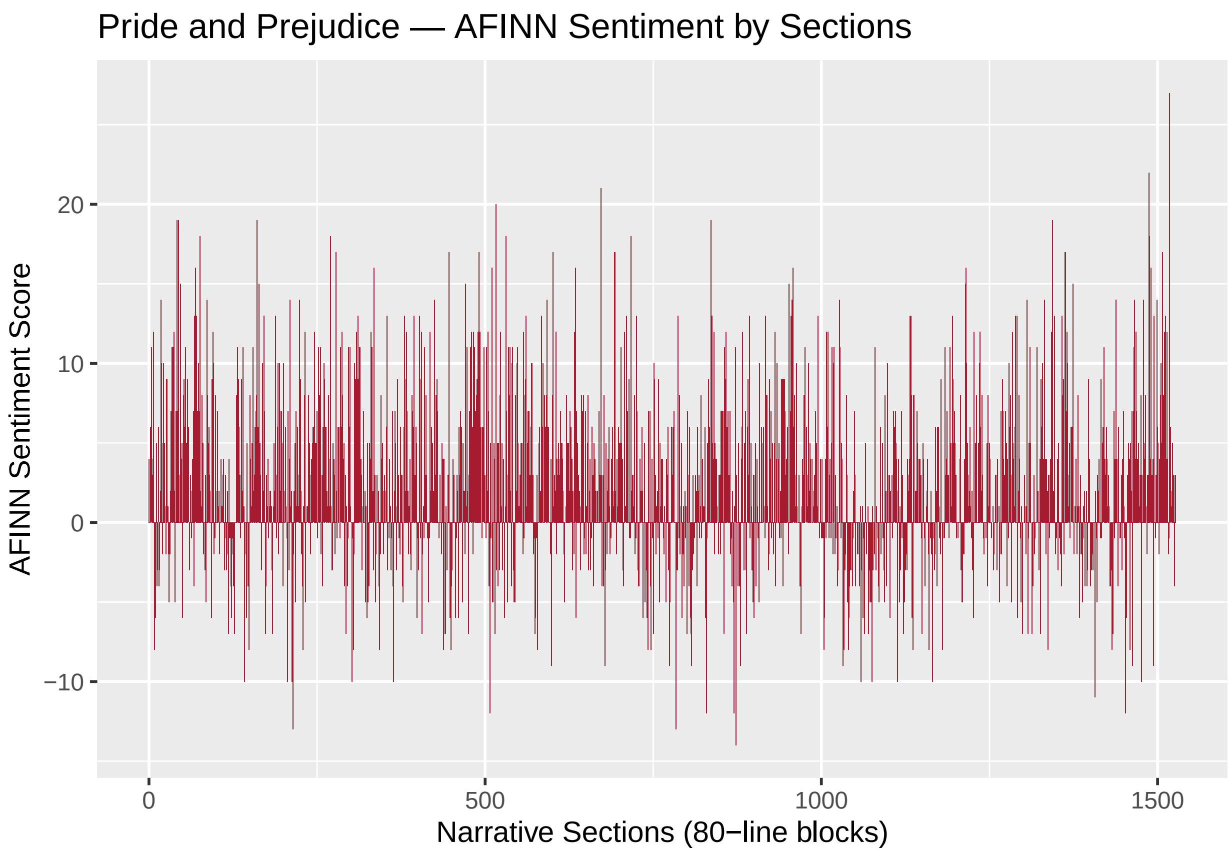
Positive and Negative Words (Bing Lexicon) Here, I counted the number of positive and negative words to understand the overall polarity of the text.

```r
pride_text %>%
  inner_join(get_sentiments("bing"), by = "word") %>%
  count(sentiment, sort = TRUE)
```

```
## # A tibble: 2 x 2
##   sentiment     n
##   <chr>     <int>
## 1 positive   5052
## 2 negative   3652
```

Sentiment Trend Over the Narrative (AFINN Lexicon) I used the AFINN lexicon to visualize how the emotional intensity changed throughout the story.

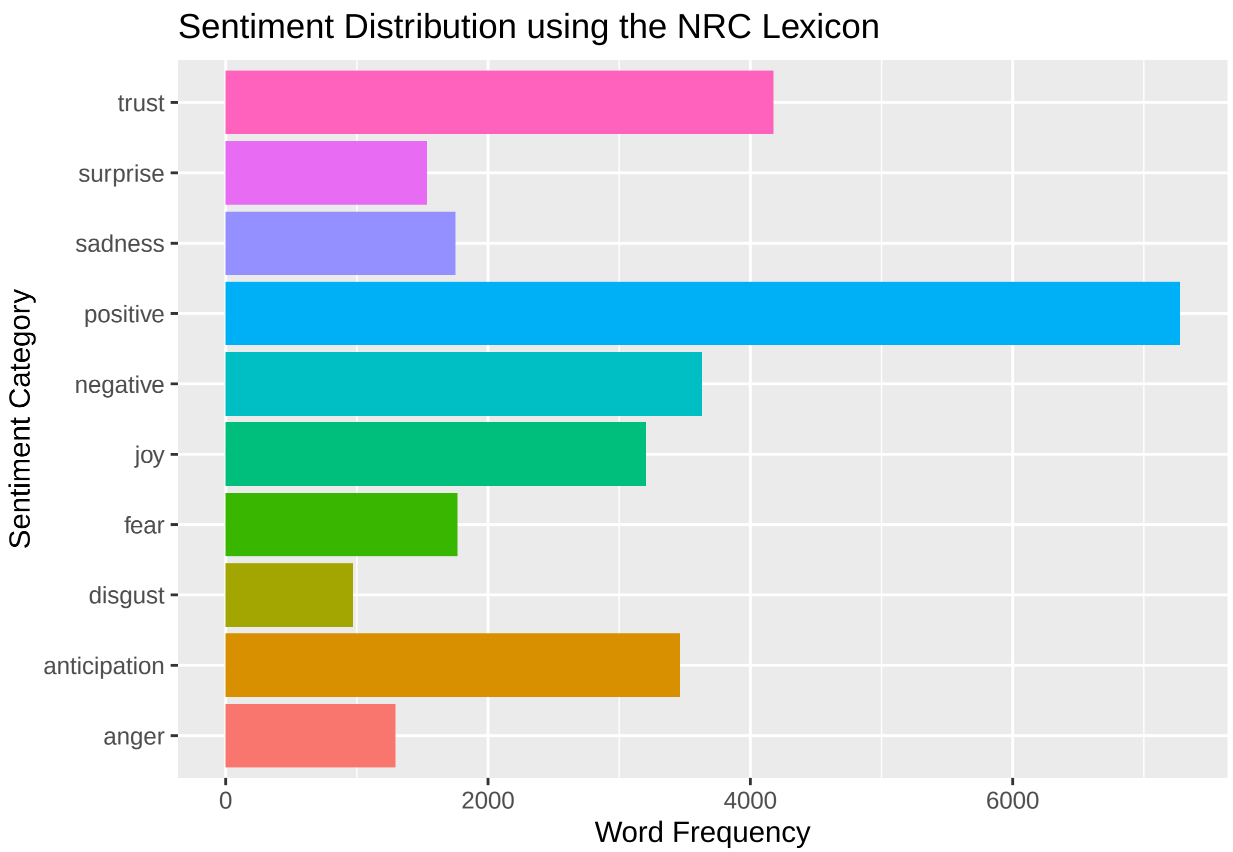
```r
pride_afinn <- pride_text %>%
  mutate(linenumber = row_number()) %>%
  inner_join(get_sentiments("afinn"), by = "word") %>%
  group_by(index = linenumber %/% 80) %>%
  summarise(sentiment = sum(value), .groups = "drop")

ggplot(pride_afinn, aes(index, sentiment)) +
  geom_col(fill = "#A51C30") +
  labs(
    title = "Pride and Prejudice - AFINN Sentiment by Sections",
    x = "Narrative Sections (80-line blocks)",
    y = "AFINN Sentiment Score"
  )
```



4. Alternnative Lexico — NRC

Finally, I applied the NRC lexicon to categorize words into specific emotions such as joy, trust, fear, and sadness. This allowed me to explore the broader emotional landscape of Pride and Prejudice.

```
sentiment_counts <- pride_text %>%
  inner_join(get_sentiments("nrc"), by = "word") %>%
  count(sentiment, sort = TRUE)

ggplot(sentiment_counts, aes(x = n, y = sentiment, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  labs(
    title = "Sentiment Distribution using the NRC Lexicon",
    x = "Word Frequency",
    y = "Sentiment Category"
  )
```



5. Summary and Interpretation

Through this analysis, I explored three different sentiment lexicons:

- Bing, which measures general polarity (positive vs. negative)
- AFINN, which captures intensity of sentiment numerically
- NRC, which classifies words into emotional categories

I found that Pride and Prejudice contains a predominantly positive tone, with strong themes of **trust**, **anticipation**, and **joy**. Each lexicon provided unique insights, helping me understand the emotional flow of the novel from different perspectives.