Sentiment Analysis

Elisa Zhang

Task one: pick a book

The book I chose is The Call of the Wild written by Jack London. I will use gutenbergr Package to download the full text and do the later analysis.

```
call_of_wild <- gutenberg_download(215)
```

1. Tidy data

First, I wrangled the book data and exclude the stop words in the book. Then I count the word frequncy and here shows the words whose frequency are larger than 50 in The Call of the Wild.

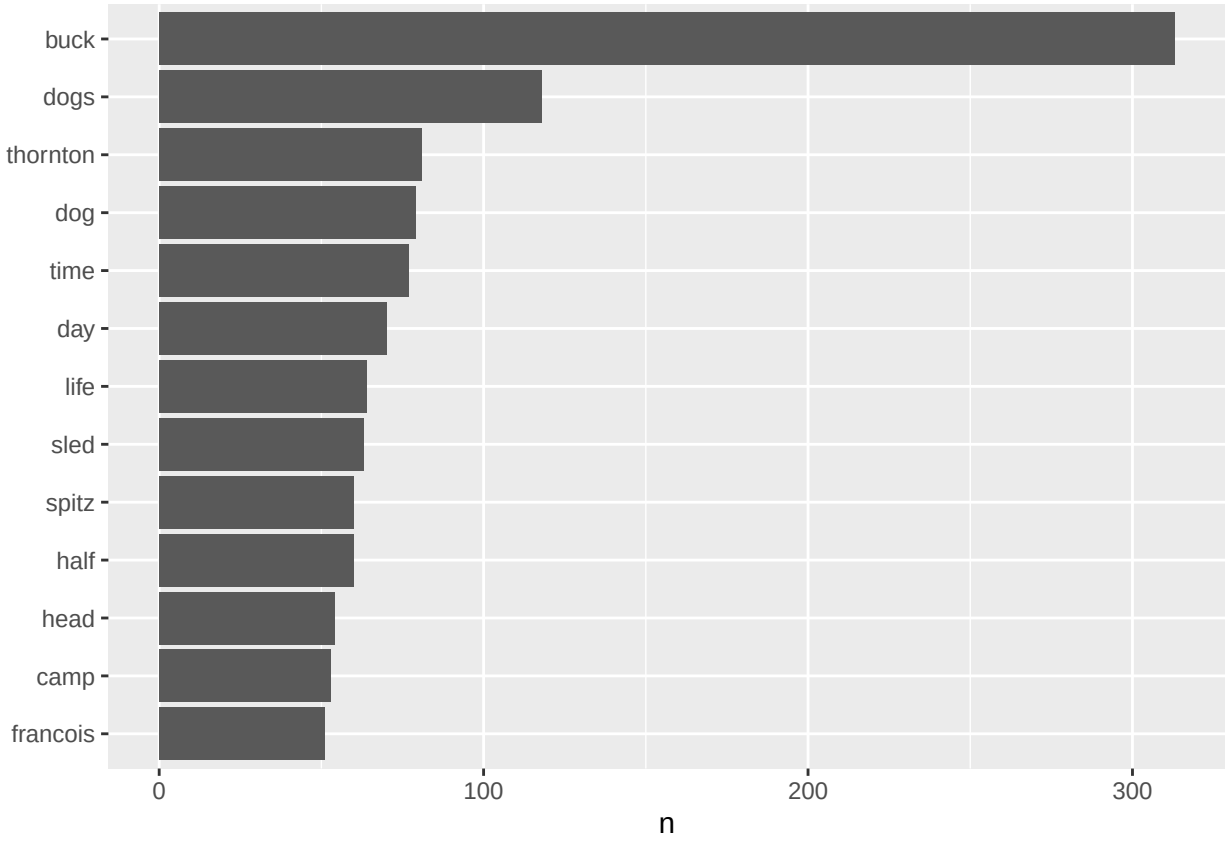


Figure 1: Word Frequncy

2.Sentiment analysis using three different lexicons

In this part, I use three different lexicons to apply the sentiment analysis on the book. When we compare three sentiment analysis, we found the result is quite different in chapter 2 where Bing lexicon only gave us negative feedback. Following the plots, I think using NRC lexicon might be better than other two. The plots in each chapter are not invariant. However, in bing lexicon, nearly all Chapter 2, 3 and 5 only show negative sentiments. And in the last chapter, Buck - the main character in the book sheds the veneer of civilization, and relies on primordial instinct and learned experience to emerge as a leader in the wild. It is obvious that the last part of the book is a mix of success and difficulties. In this way, NRC lexicon is the best in sentiment analysis among three lexicons.

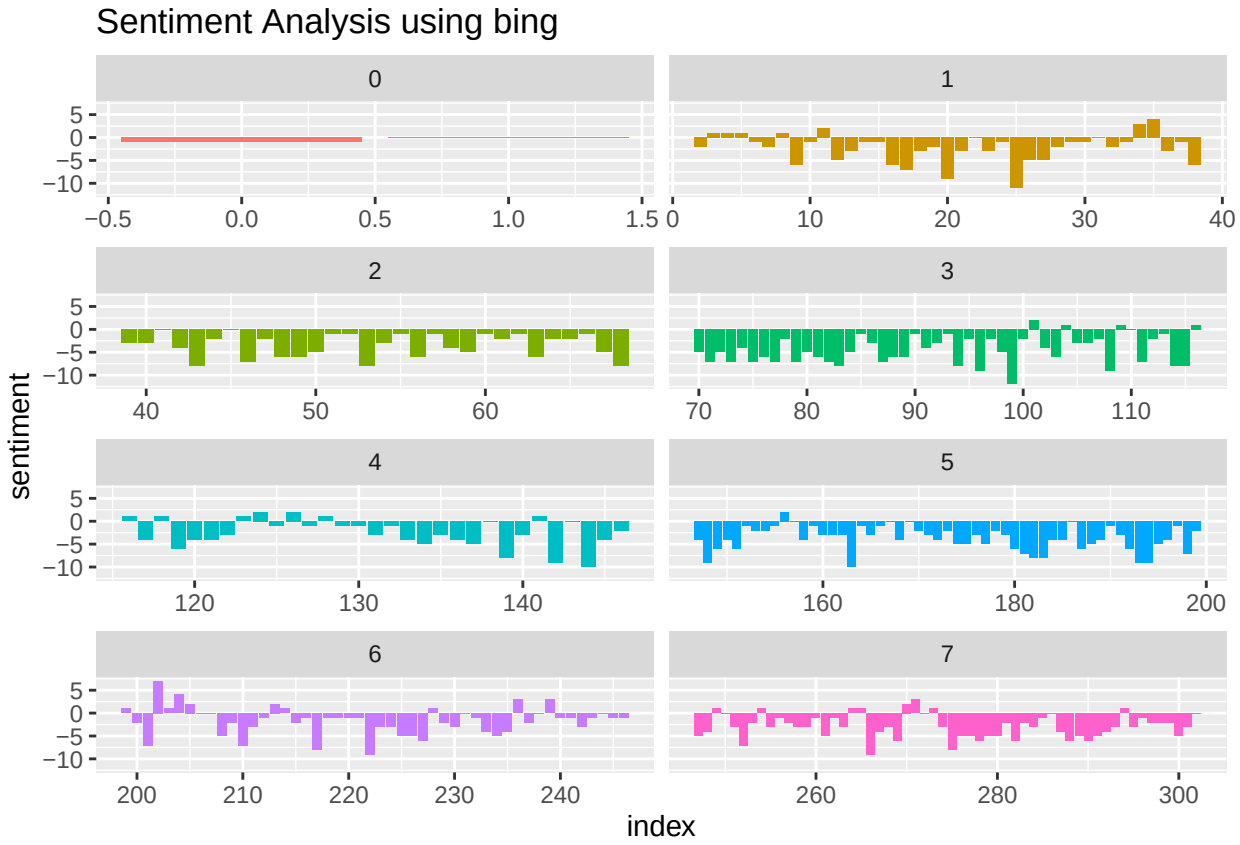


Figure 2: Sentiment Analysis differnet lexicons

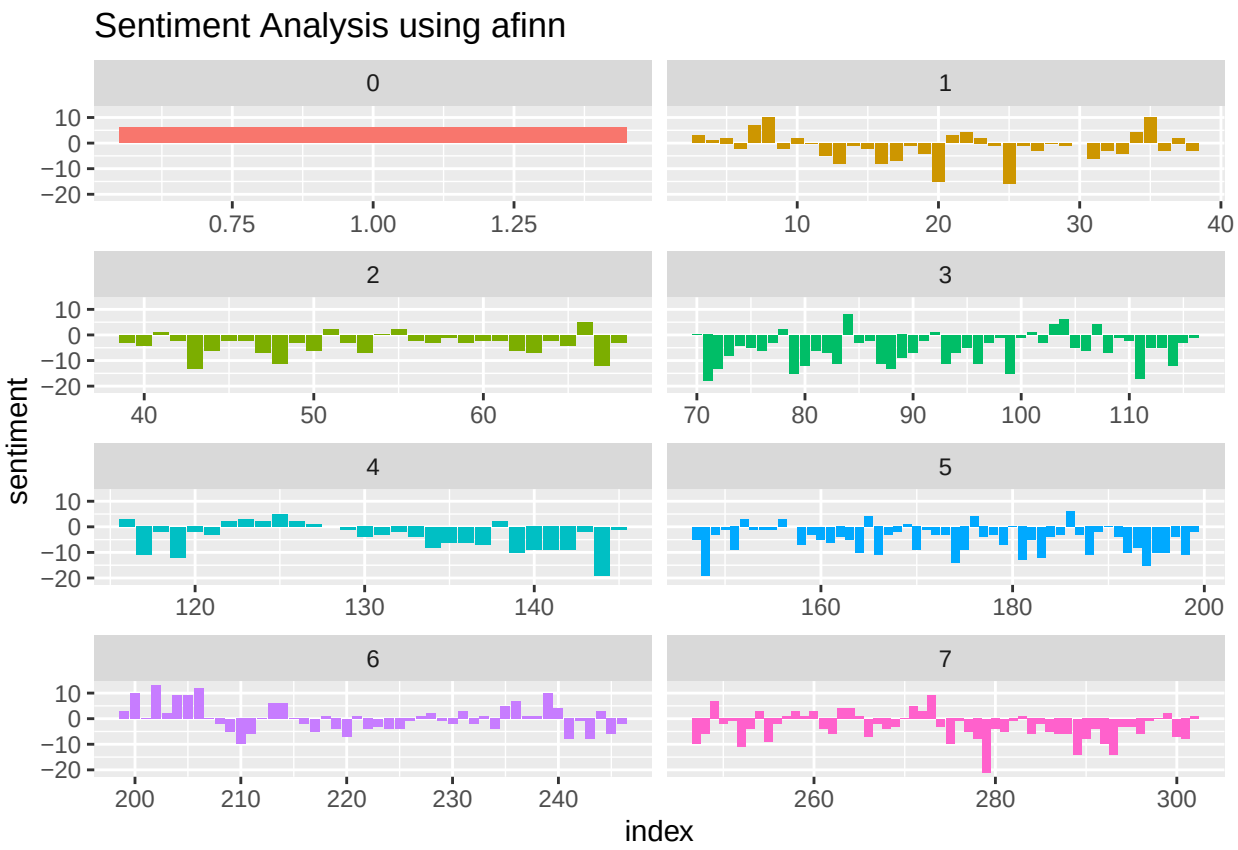


Figure 3: Sentiment Analysis differnet lexicons

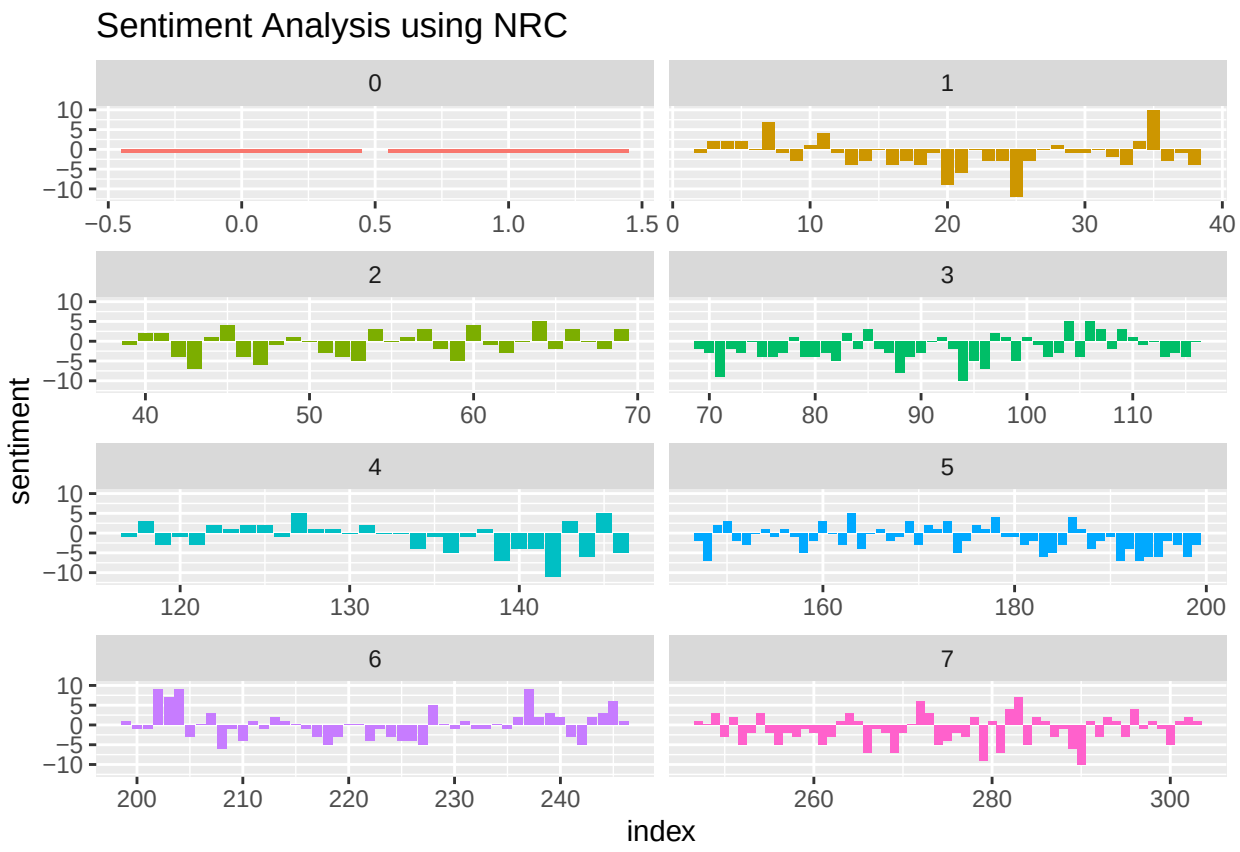


Figure 4: Sentiment Analysis differnet lexicons

3. Most common positive and negative words in The Call of the Wild

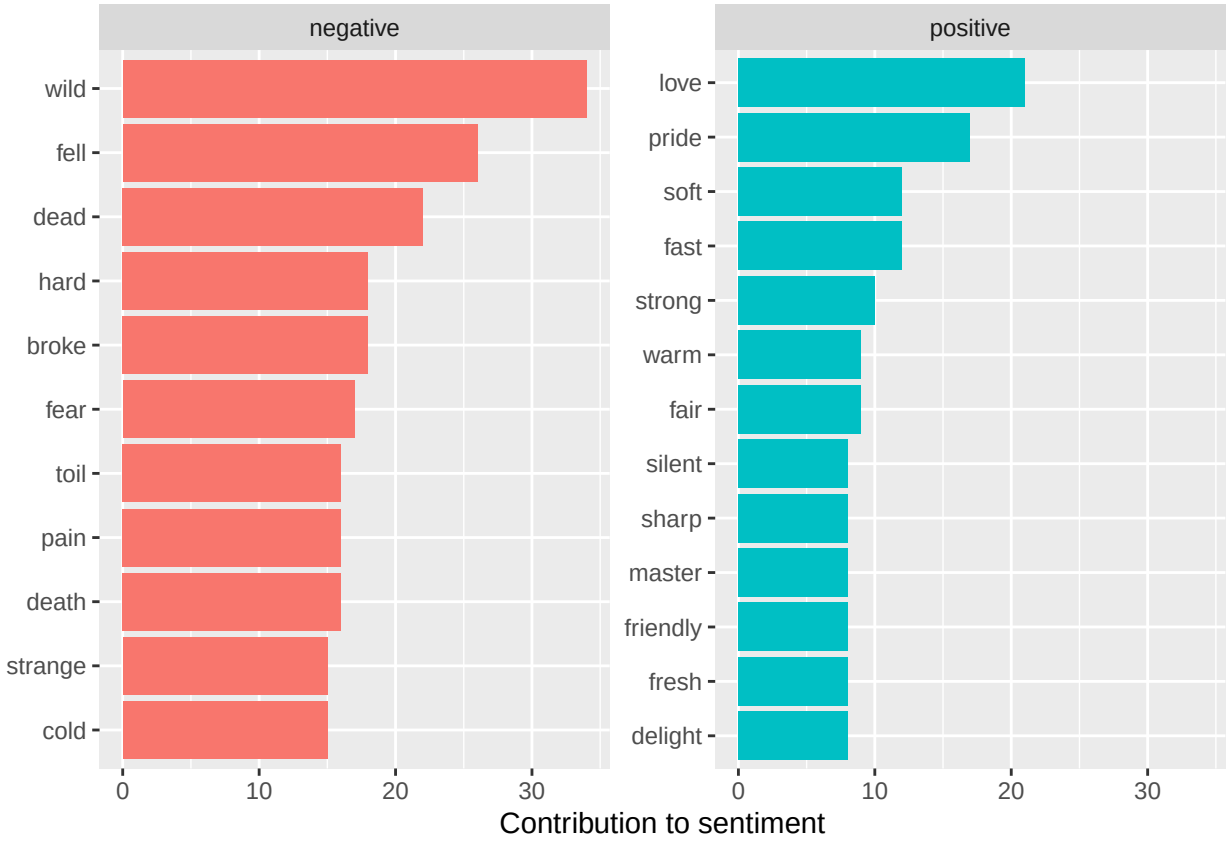


Figure 5: Contribution to the sentiment using Bing

We analyze word counts that contribute to each sentiment. When we using Bing lexicon, the top 1 word that contributed to negative sentiment is wild. The analysis is not precise. For Buck - a dog, wild is his final and best home. However, in the positive part, the result is reasonable.

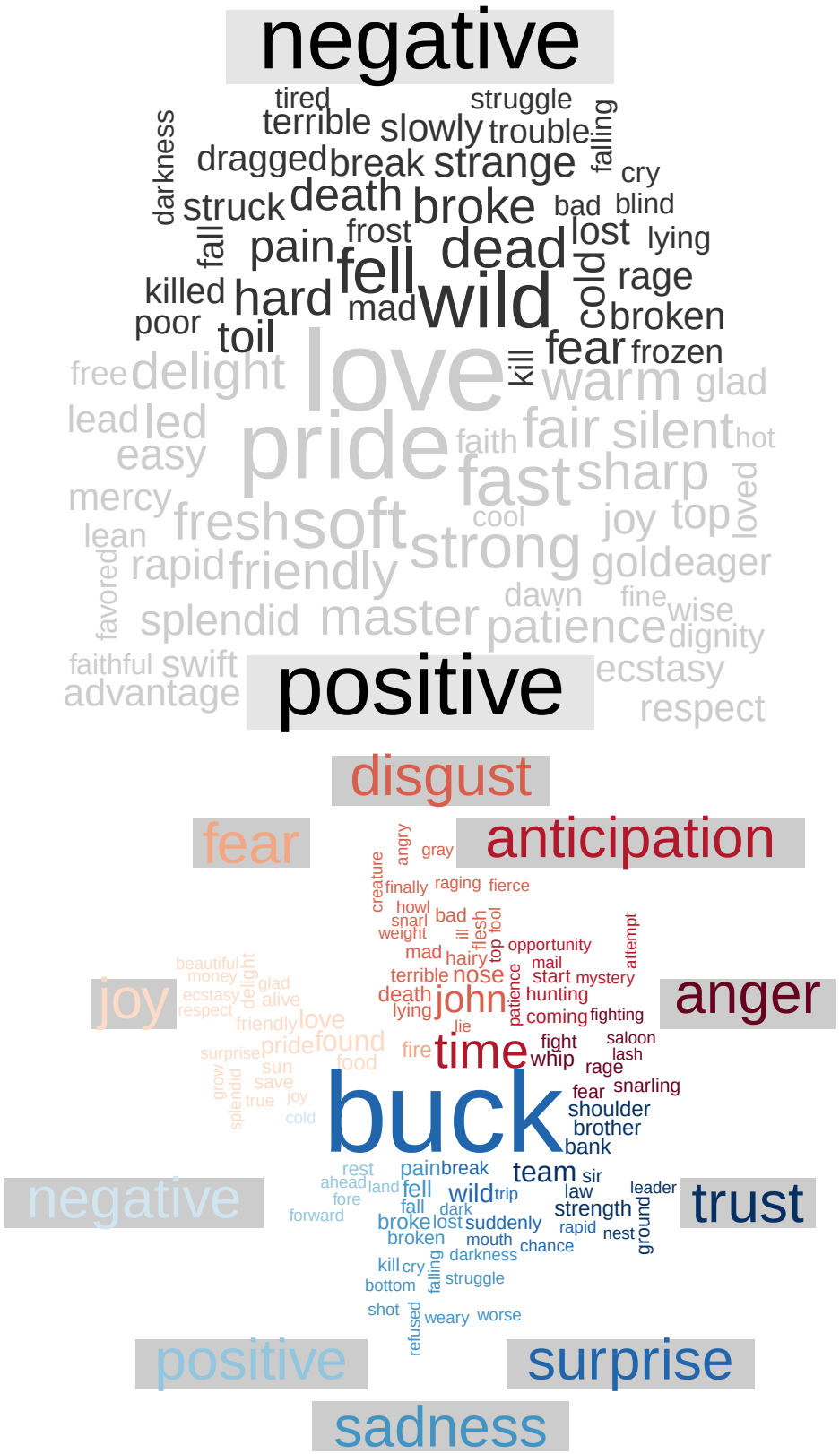
When first implement the analysis using nrc lexicon, there is an anomaly: the word 'buck' is the name of the main character. We add the name buck to the stop_words and re-run the analysis.

When compared the most common positive and negative words in the book, I used two different lexicons. I might not tell which one is better. They all have strengths and weakness. It might indicate that it is not enough to implement word level sentiment analysis.

Figure 6: Contribution to the sentiment using NRC

4.Wordclouds

Here we show the most common negative and positive words using bing lexicon in the book.



5. Additional Lexicons

I will use an additional lexicons called loughran which is created from financial report.

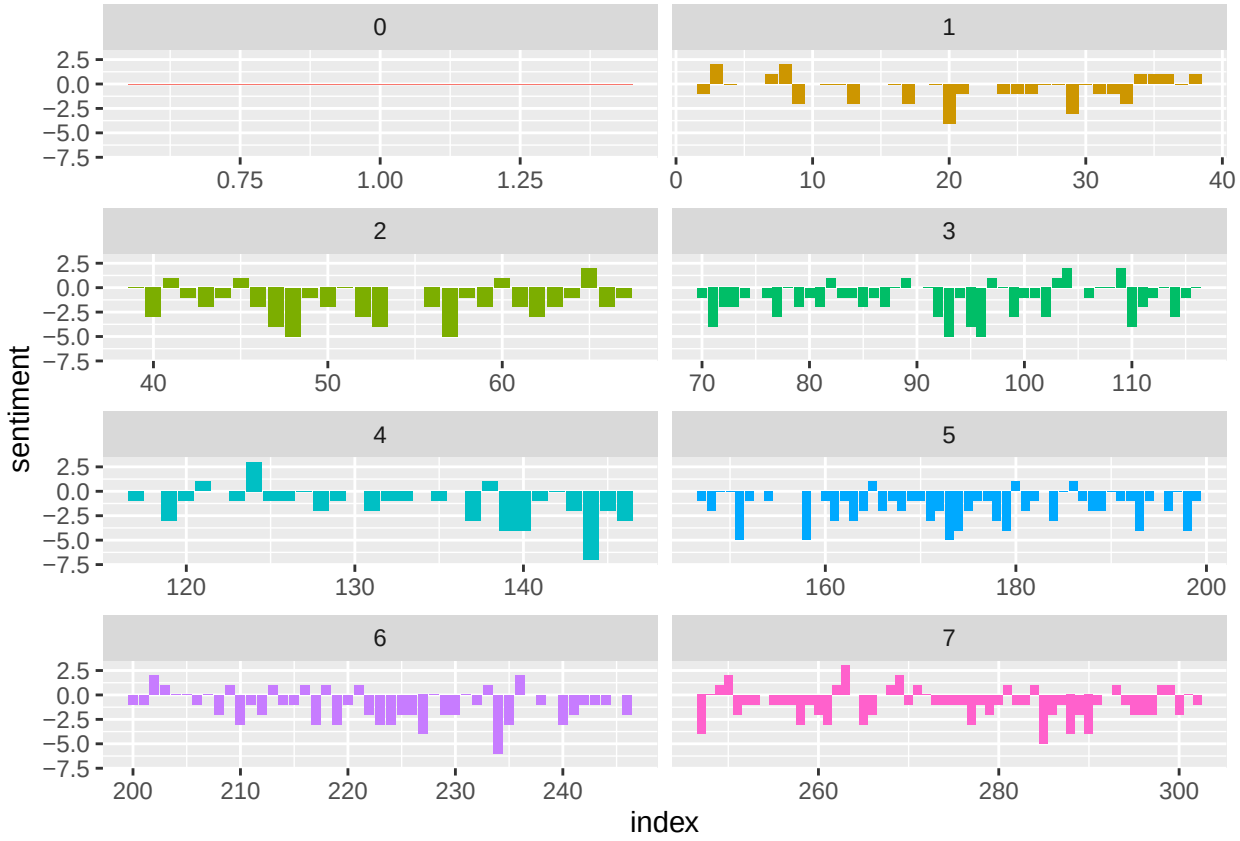


Figure 7: Sentiment Analysis using loughran

Compare sentiment analysis using 4 types of lexicons.

Since loughran lexicon is designed for financial report, its performance is not good as the language is quite different between short fictions and reports.

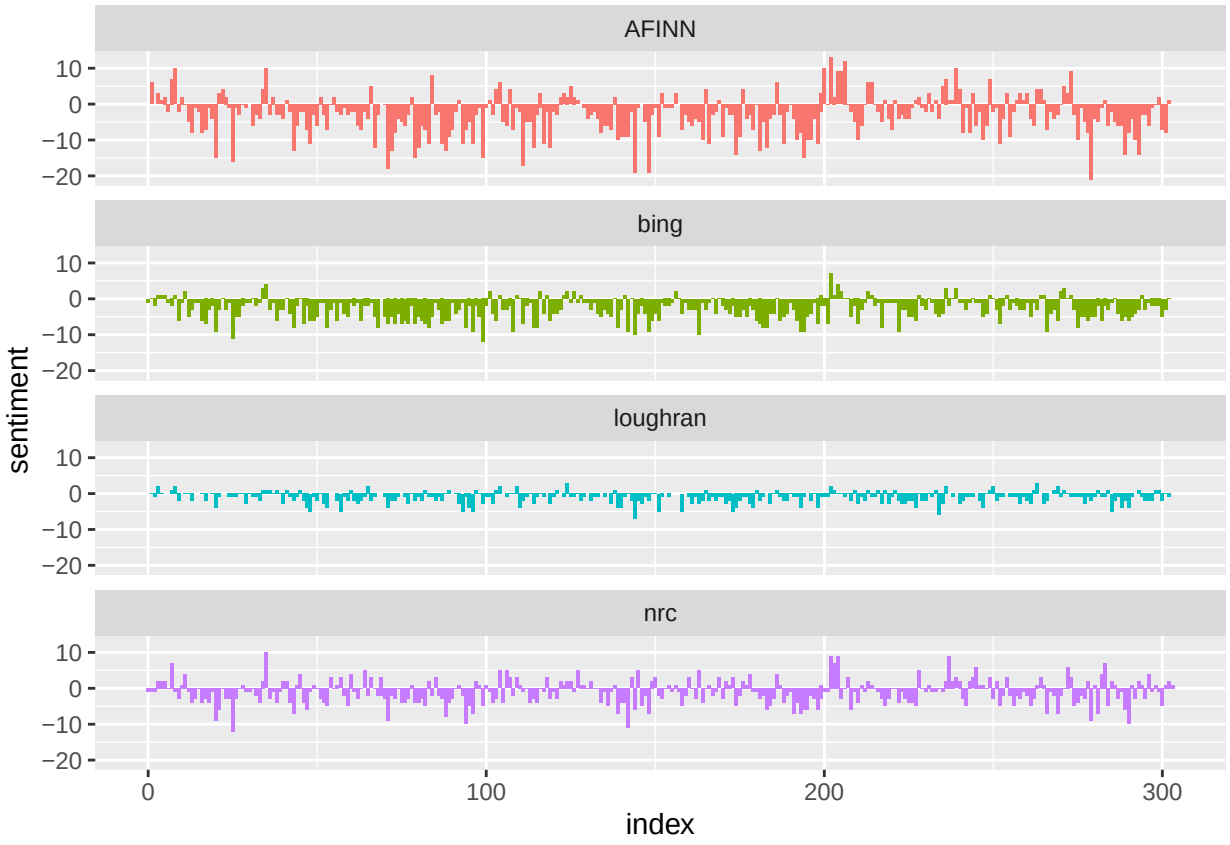


Figure 8: Comparison 4 types of lexicons