MA615 Assignment 4: Text Analysis

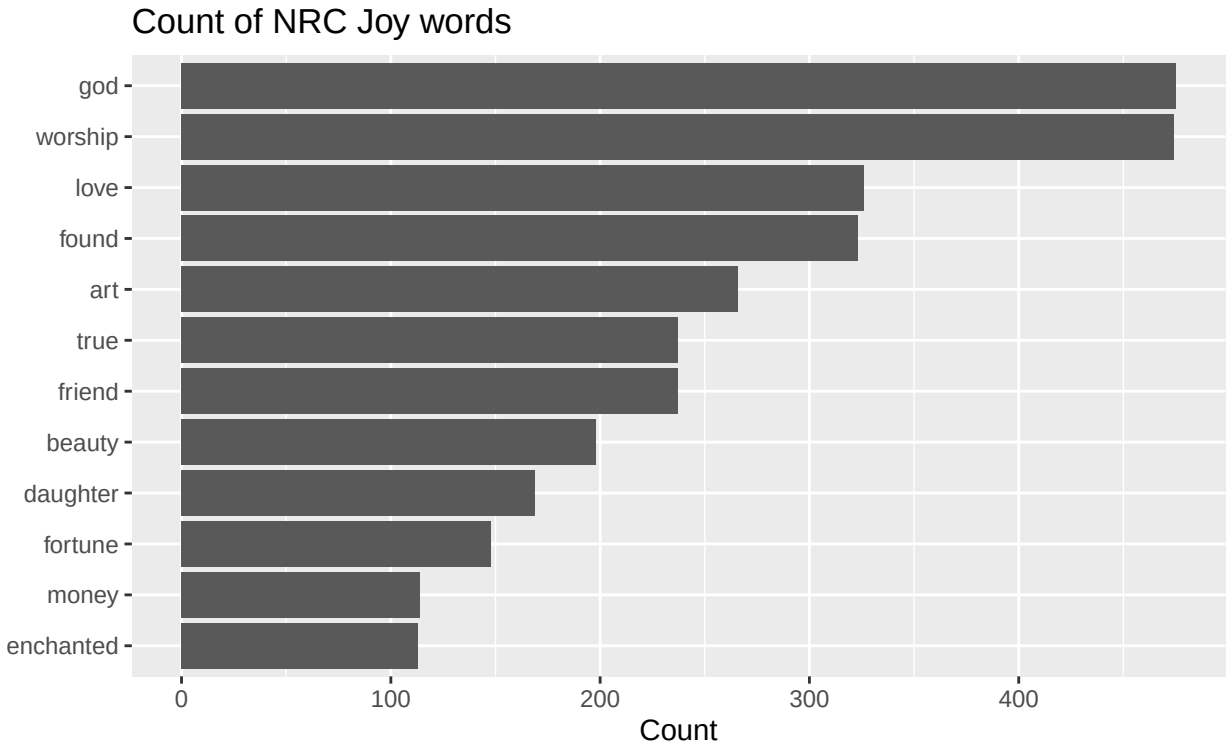
Jack Carbaugh

12/6/2021

Don Quixote Bag of Words Analysis

For this project, I will be analyzing Miguel de Cervantes' Don Quixote. The novel is quite long, and is well known for being both comedic and deep in many facets. As a general overview, the novel follows Don Quixote, a crazed elderly man who thinks he's a knight, and goes about the Spanish medieval countryside causing mayhem. His witless companion Sancho Panza follows him as his squire, and faithfully goes along with his master's antics. The story is mainly meant to be humoruous, while also exploring the relationship of these two characters, and the psychology of the people they meet along the way.

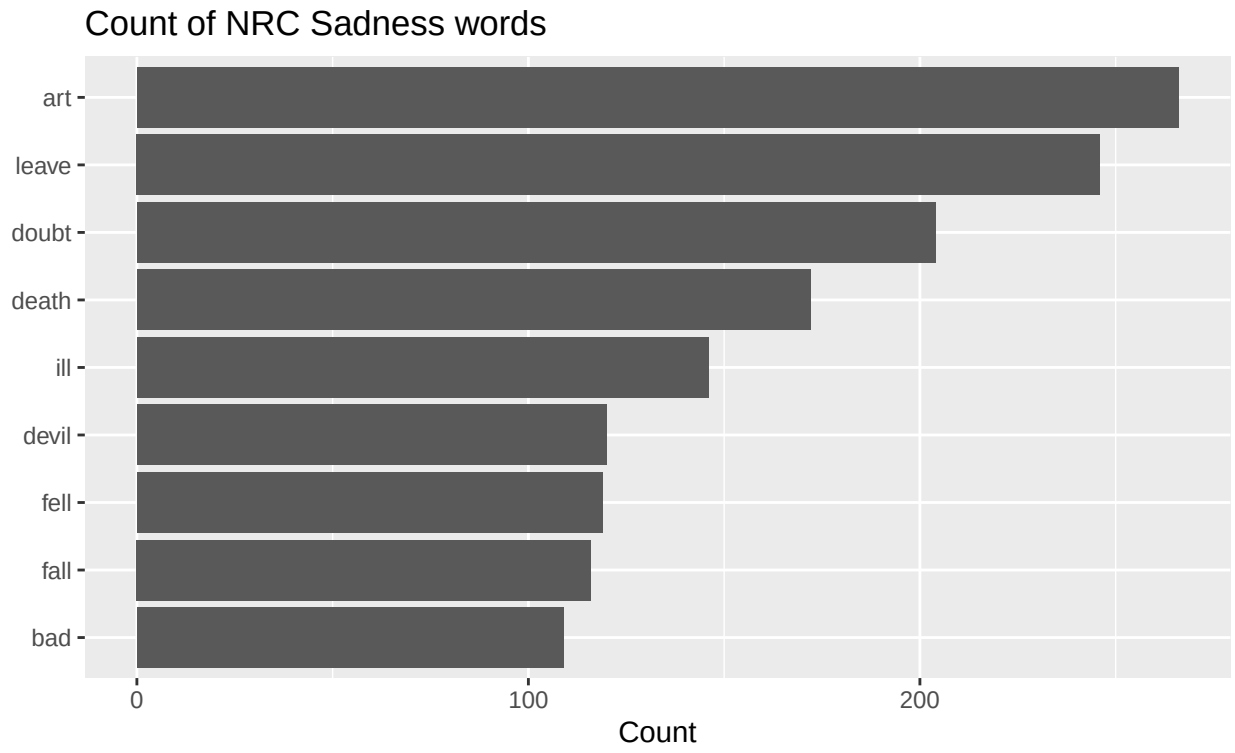
For the sentiment analysis of Don Quixote, I first wanted to get a sense of the type of words used in the work. After downloading the book using the gutenbergr package, I Used the tidytext package to break the book down into words, and then remove all stop words like "the", "of", "and", etc. This left me with a dataframe where each row contained a word in the book, the line number it appeared in, and the chapter it appeared. To do an initial check of the senitment, I used the NRC lexicon, which contains a bunch of words labeled by 10 different sentiments, depending on whether or not they match the sentiment (according to NRC). Looking at the sentiment of Joy in Do Quixote, these were the words that appeared with high frequency:



Here we see that some of the most joyful words that appear relate to religion and love, common themes of

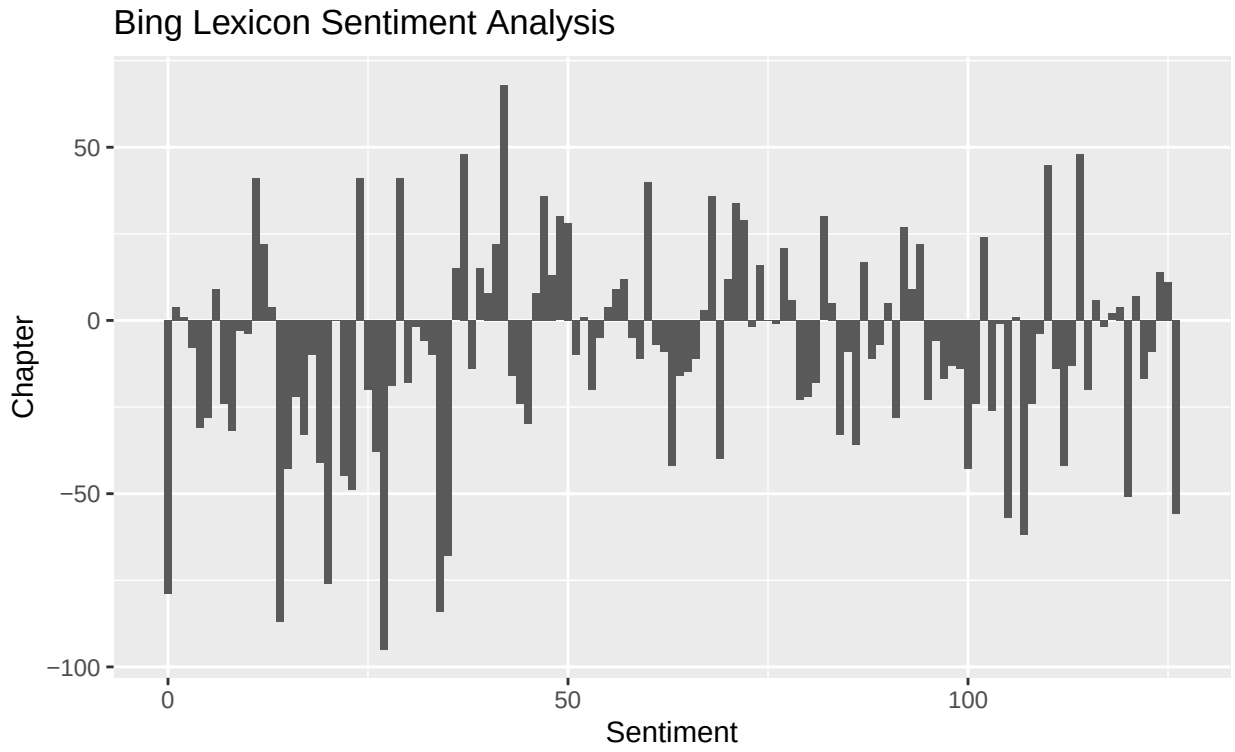
the book. “Found” is also common, but may not be considered joyful in the context of the novel.

We can do a similar analysis for the opposite emotion of sadness.



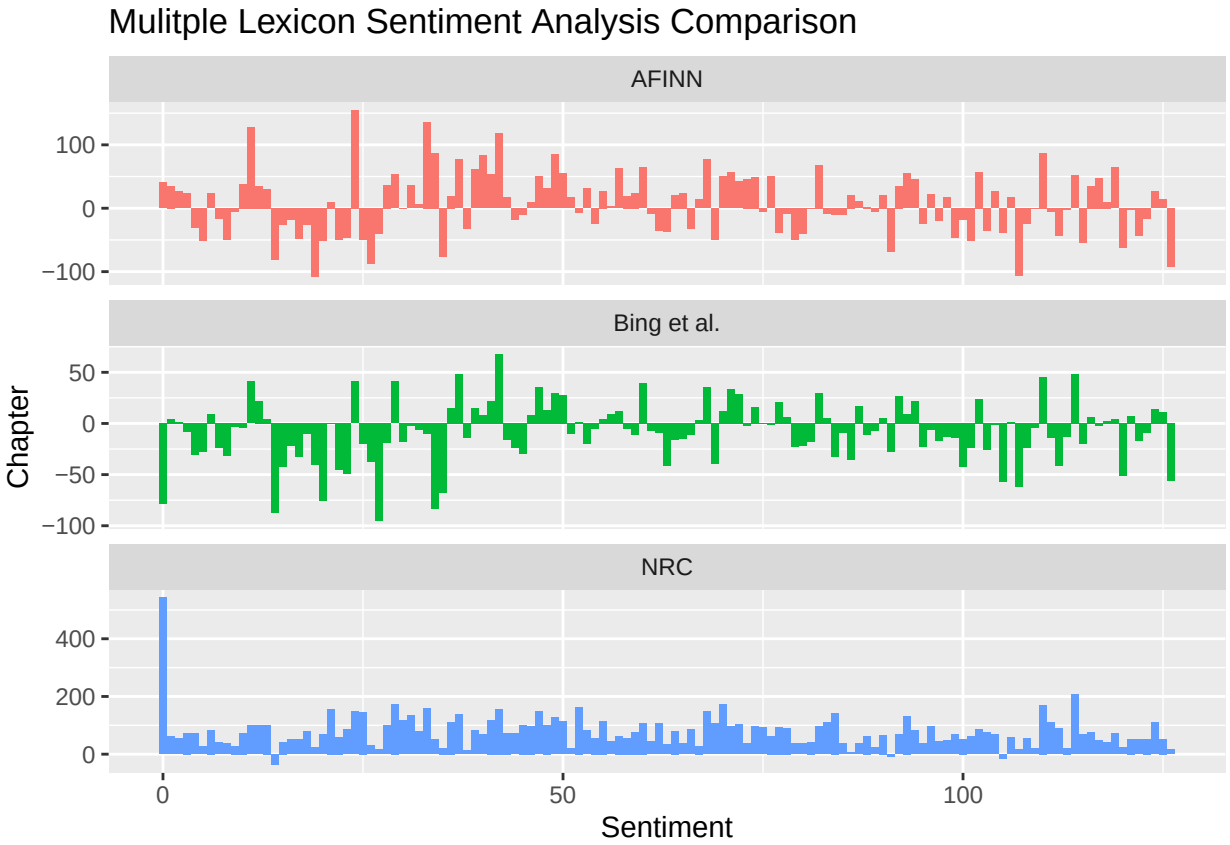
Here we see that some of the most common sad words are “leave”, “doubt”, and “death”. This also fitting to the text, as Sancho is frequently frightened by Don Quixote leaving him alone, often doubts Don Quixote’s judgement, and is fearful of his death. An oddity of this lexicon is art being the highest frequency sad word, while it also appears in the joy category. This may skew our analysis, as art is generally looked upon happily in the novel.

Next, we'll want to see the progression of the sentiment of the novel as the plotline flows. I first created a new dataset using the Bing lexicon, which only rates words in the two categories of positive and negative. I then counted up the number of positive and negative words in each chapter, and found the difference in these counts as the sentiment. The book is long with many short chapters focused on a particular issue, so I felt splitting over chapter made logical sense.



This follows along with the plot of the book fairly well. There are clearly many more negative spikes then positive spikes, aligning with the frequency of the negativity and dark humor throughout the novel. Don Quixote and Sancho consistently meet violence and trouble, especially, within the first 30 chapters of the book, so this aligns well with the analysis. Chapters 30 - 50 are generally more positive as we follow the story of Dorotea and Camacho. Their story is tragic at first, but ends with happiness for many parties around the 50 chapter mark. Chapters 50 - 70 follow Don Quixote and Sancho creating trouble in the countryside once again leading to the greater negativity. Chapters 70 - 90 invovles Don Quixote and Sancho in the Duke's castle, which has some high moments for the pair, but also many lows. Finally, the last chapters generally involve Don Quixote becoming increasingly downtrodden, losing a fight against another knight, and ultimately dying, Explaining much of the negativity in the end.

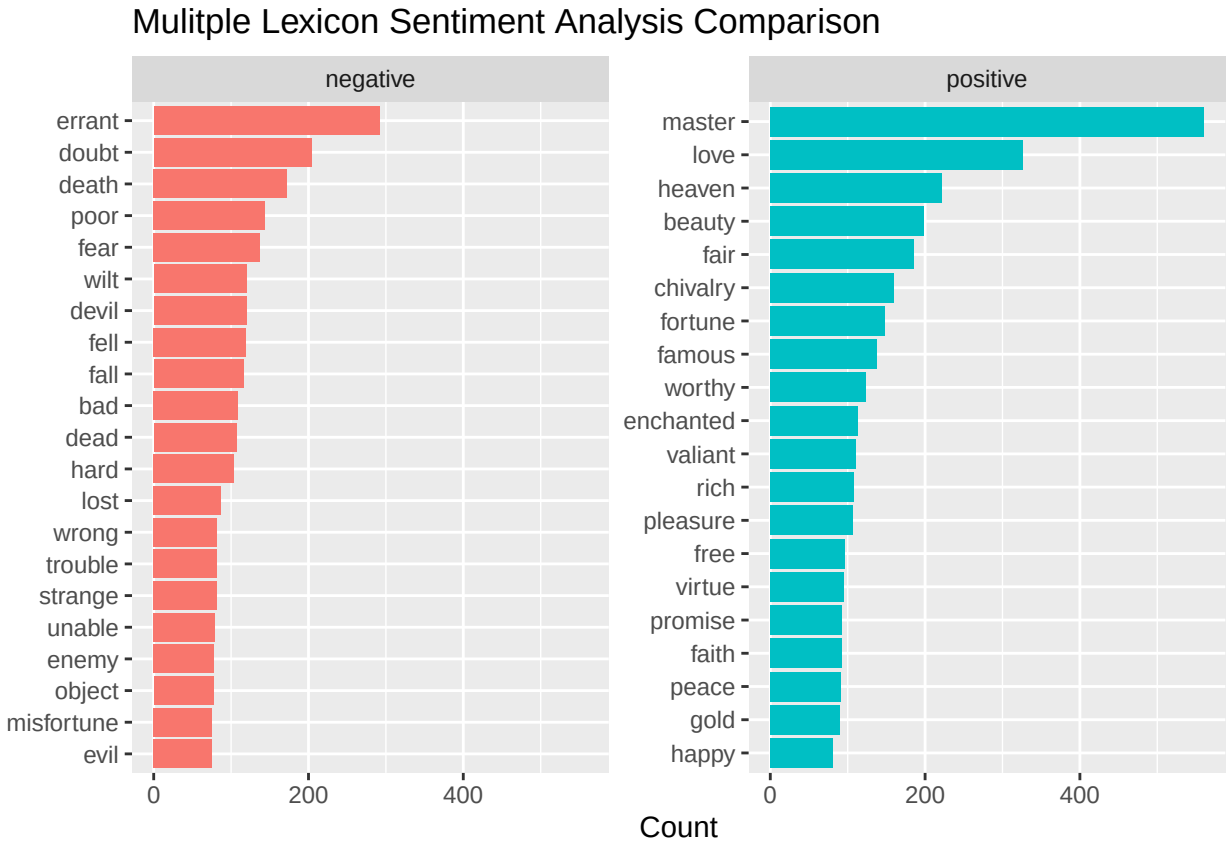
However, this is only one lexicon. It will be helpful to compare it to others: the NRC and Afinn lexicons available from tidytext.



The Afinn lexicon rates every word on the integer scale of -5 through +5, where the higher positive number means higher positivity, and the lower negative number means higher negativity. The dataframe was made by cross referencing this lexicon to find the value of each word in Don Quixote, and then grouping and summing up the values by chapter. For the NRC portion, I singled out the positive and negative sentiments, and found the difference in counts of positive and negative words to get a sentiment, just as was done with the Bing lexicon. These three dataframes were then rowbinded.

We see that the Afinn and Bing analyses match up very closely, with the Afinn only being slightly scaled more positively all around. The NRC analysis is a bit different, as each chapter is almost always rated as being postive overall. While on much smoother scale though, we can still vaguely see that peaks and valleys that align with the other two lexicons.

Finally, we can take a deeper dive into the Bing analysis to see which words had the greatest impact on the positive and negative scores of each chapter.



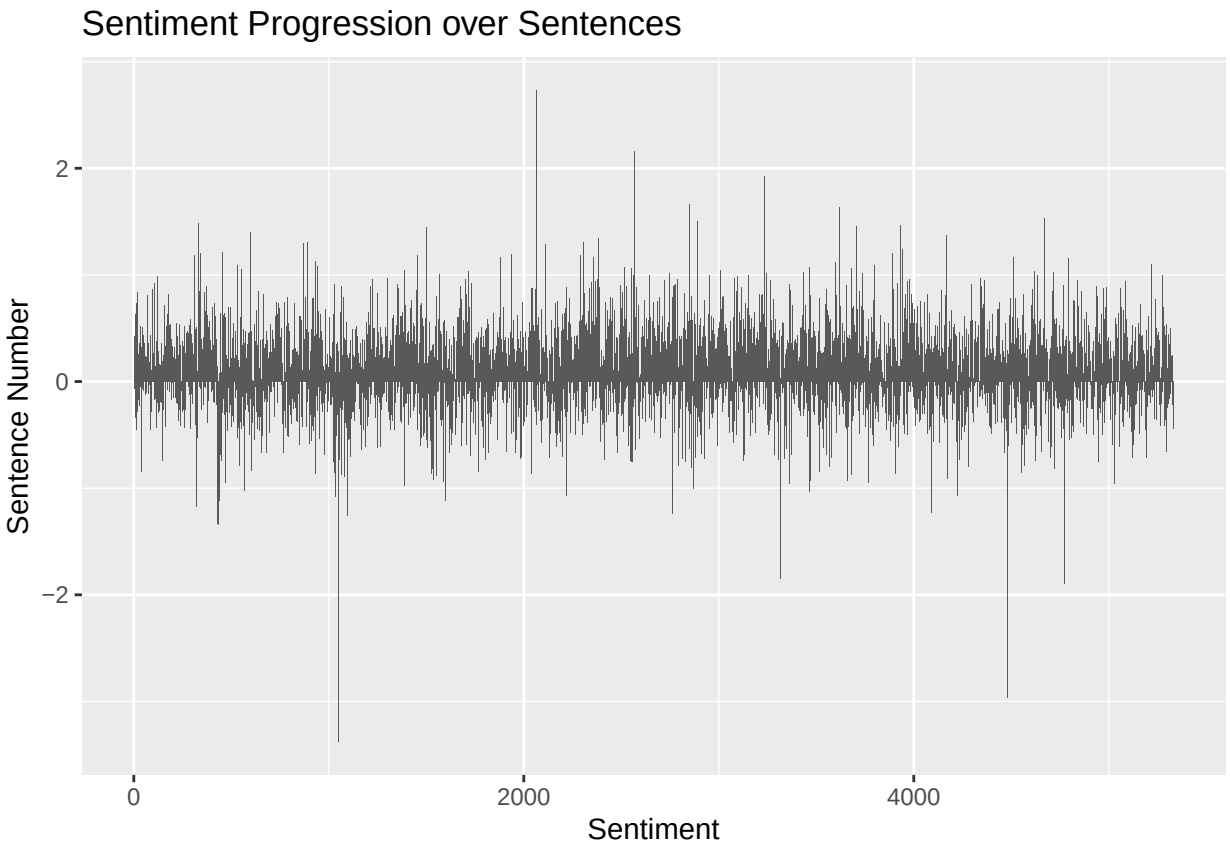
Many of the words from the first Nrc Joy analysis appear once again. However, some outliers are quite noticeable in both sentiments. The highest frequency negative word is "errant," however this word is almost exclusively used in the phrase "knight-errant," the lifestyle Don Quixote considers himself to be leading. Thus, the term is not truly used in a negative context, and may be considered to remove. Another oddity is the great frequency of "master" in the positive section. This is likely due to Sancho frequently referring to Don Quixote as his Master. Just like with "errant," the word is more used as a neutral noun, but Sancho does occasionally refer to Don Quixote affectionately as his "master," so it may suffice to keep it in. Finally, an interesting word in the positive section is "enchanted." While normally positive term, Don Quixote is concerned with his love Dulcinea being transformed into another woman by an enchantment. As Don Quixote is quite distraught over this, it may make sense to alter the sentiment of "enchanted" to negative.

Sentence Level Analysis

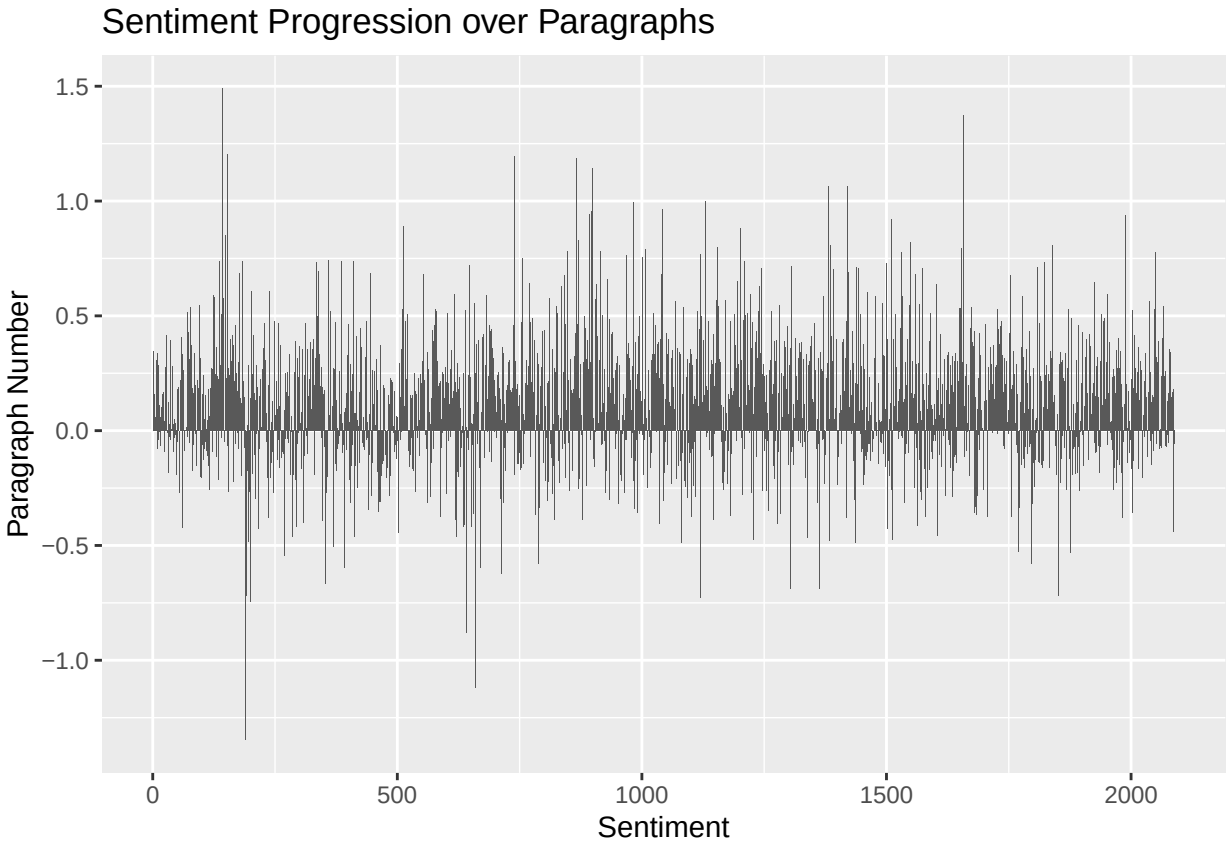
Using the tnum library, I ingested a slightly altered version of the Don Quixote text. True numbers was able to successfully break up the book into sentences, and then label each sentence by the chapter and paragraph it is in. Using a query, I filtered out all of the text from the true numbers database.

Now, I can utilize the sentimentr package to help with sentiment analysis, as this package can determine the sentiment of a sentence. It does this using a polarity algorithm, and finds scores for individual words, as well as if those words have been negated using terms like "not." In the end, each sentence can be given a numerical sentiment score, where greater positive values indicate higher positivity sentiment, and lower negative values indicate lower negativity sentiment.

First, I used the sentiment function to simply get the sentiment of each sentence in the book, and then plotted a column chart.

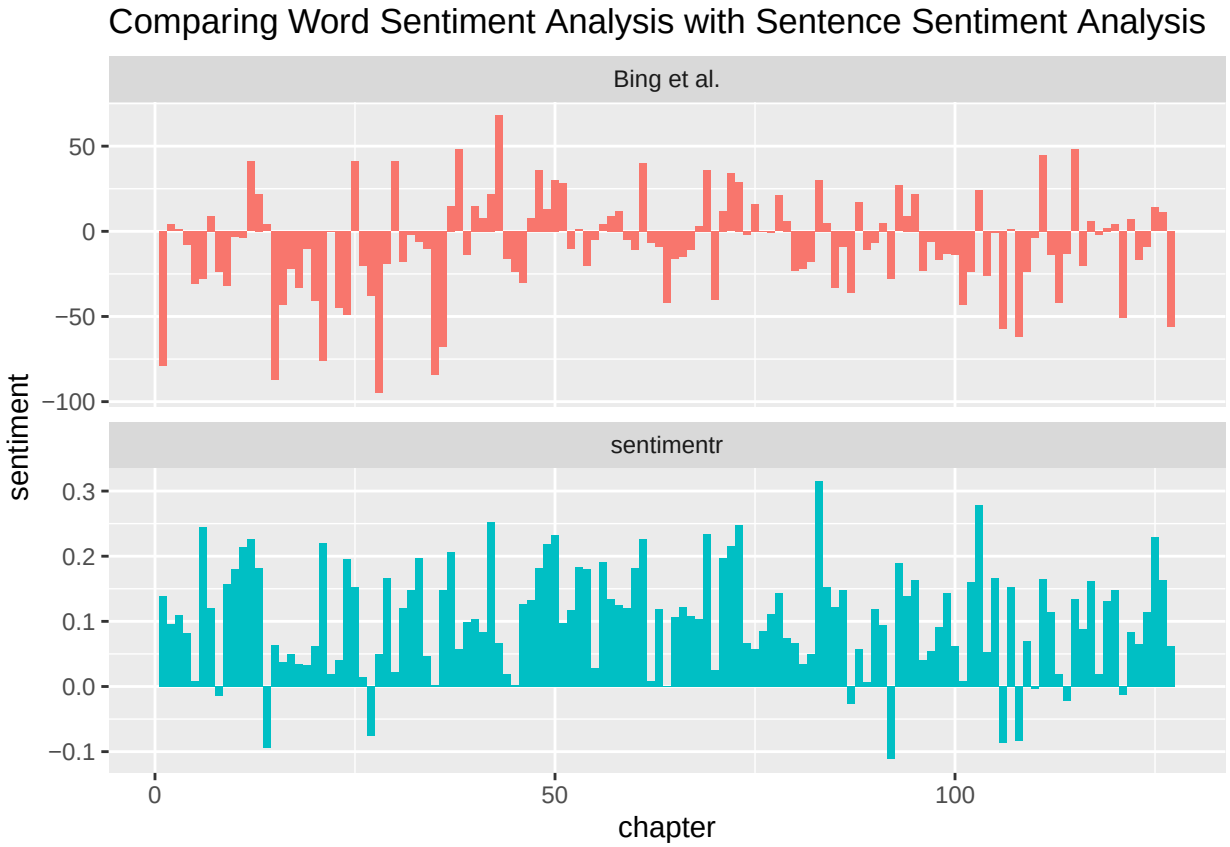


However, as there are such extraordinary amount of sentences in this long novel, I decided to average the sentence scores over paragraph. This was done by separating out the query root, and creating and index column for paragraph and chapter. We can now do a similar analysis over each paragraph



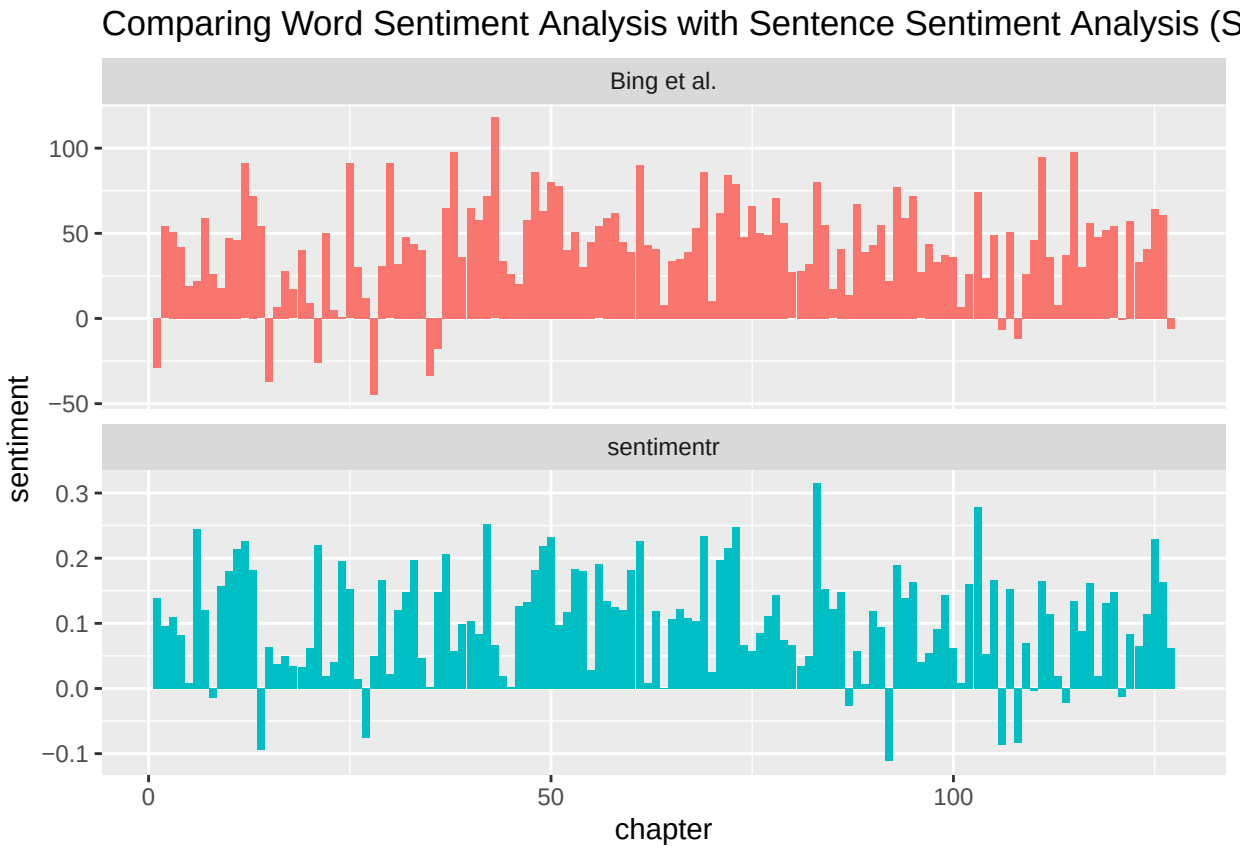
Now the progression is a bit more clear. In general, we see that paragraphs are more likley to be positive then negative overall. This lexicon seems to rate a majority of Don Quixote's paragraphs between a sentiment score of 0 and 0.3, with occasional negativity and positivity spikes aligning with the chaotic nature of Don Quixote's adventures, and he can go from positivity to negativity in the blink of and eye.

Next, I would like to compare the analysis the sentimentr produces to the Bing lexicon that was used in the previous sections. This can be done by averaging over chapter for the sentimentr analysis, and then joining the rows of thr Bing analysis into one dataframe.

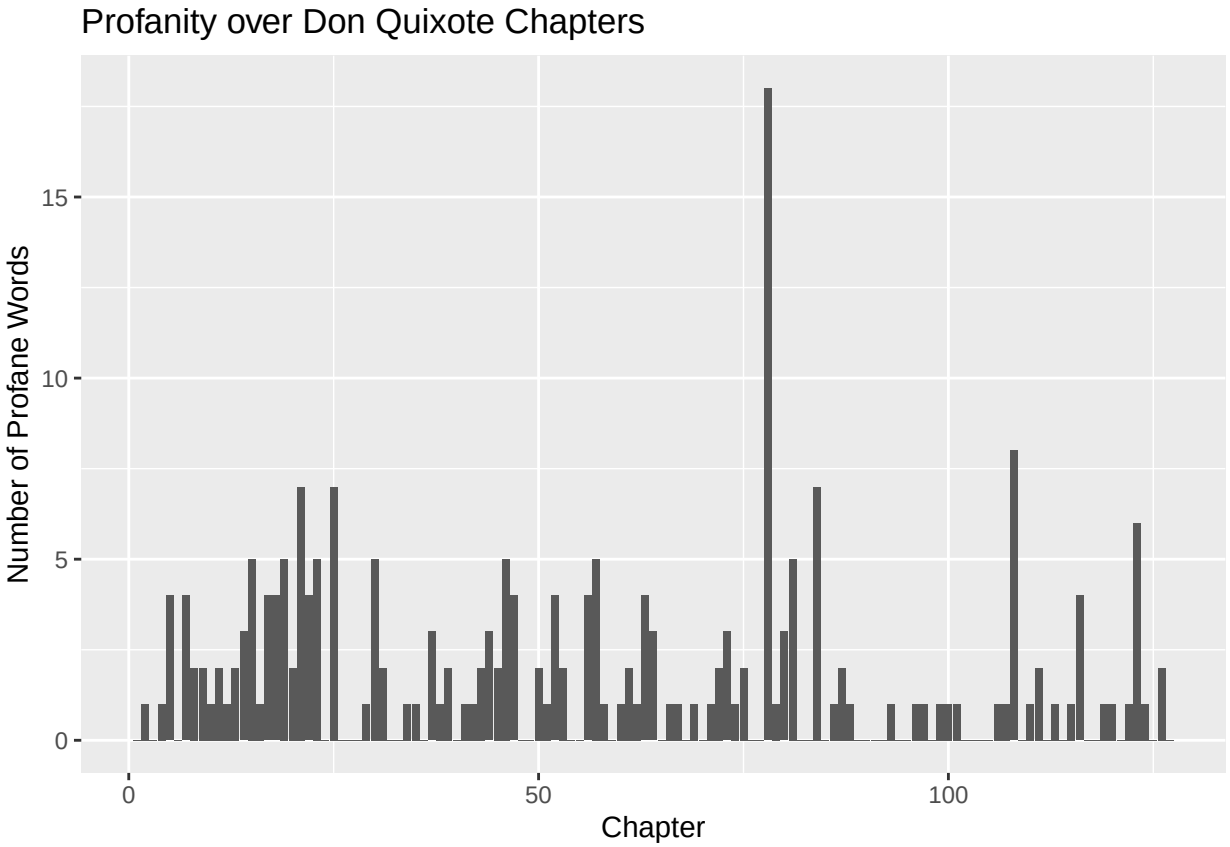


Here, we see that the most notable difference is the high positivity of sentimentr compared to the greater negativity of Bing. Regardless, similar trends are still seen despite the scale difference. In particular, chapters 12-32 both show a long stretch of mostly lower senttiment, with a few positive spikes. Then, chapters 33 - 60 stay as generally the most positive parts of the book. The only area which looks a bit off is chapters 75-100, as the Bing lexicon sees it more positively overall than sentimentr. In general though, the two books seem to agree rather well, aside from the scaling difference.

To more easily see this analysis in change, we can scale the Bing progression by adding a flat 50 to every score. With this, the similar trends of both analyses is much clearer

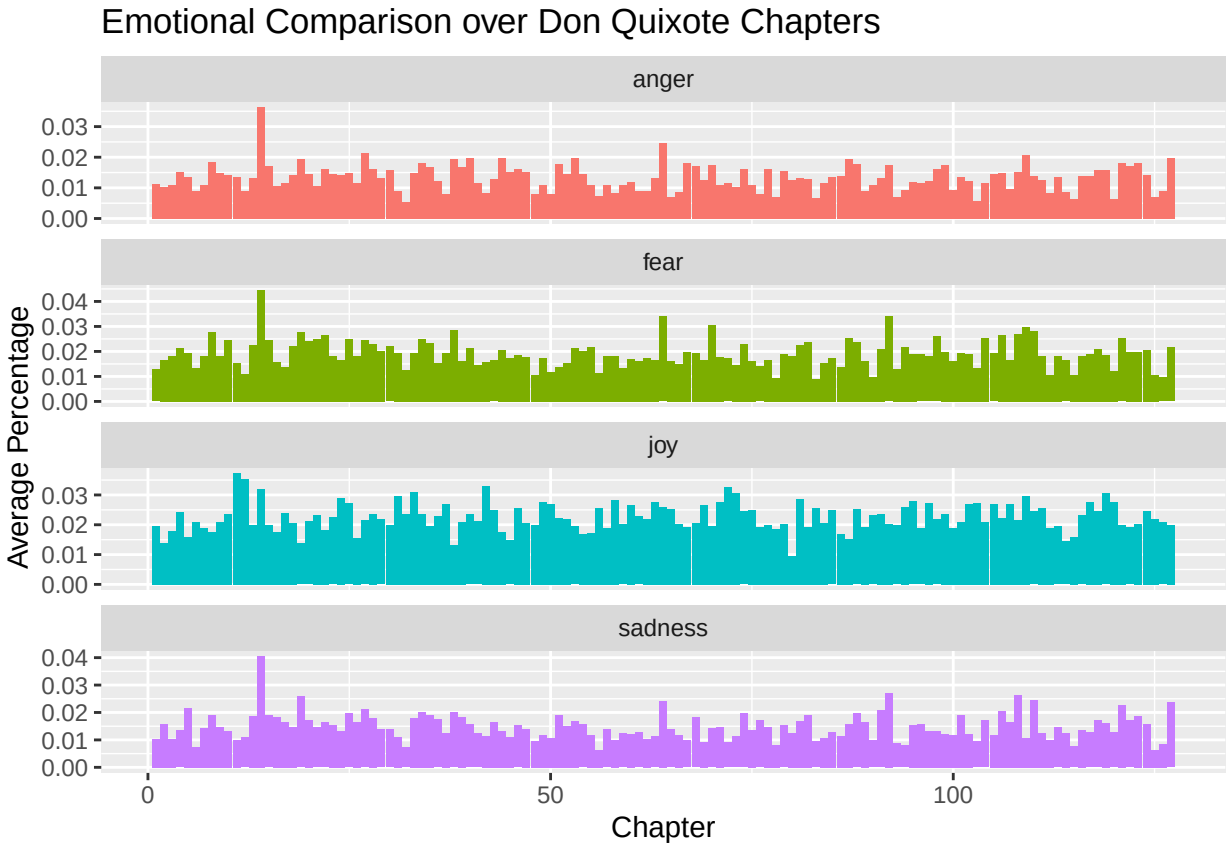


For some final analysis, we can utilize the other lexicon-based analyses in sentimentr. First, we can observe the profanity within Don Quixote. The profanity function utilizes the profanity_alvarez lexicon, and simply counts up the amount of profane words. For more clear analysis, we can observe break down by chapter again.



Looking at the y-axis, we see that overall the frequency of profanity in Don Quixote is very low. This is as expected, as many of the profane words in this lexicon are more modern, while the novel is around 400 years old. The interesting is the spike at chapter 78. Looking directly at the text, we see this spike is almost certainly due to the frequent use of the word 'ass' in this section, referring to the type of mule that Don Quixote encounters in this chapter. Thus, even this spike is simply due to the change of use of words, and not so much profanity,

Finally, we can observe some of the other emotional frequencies in Don Quixote using the emotion function from sentimentr. This function counts up the number of words used in 8 emotional categories, as well as when they are negated. To simplify, we can focus in on Anger, Fear, Joy, and Sadness



Overall, we can see that Joy and Fear are emotions that appear more frequently than Anger or Sadness. The progression of emotions over the plot of the book is also interesting. Anger and Joy seem like fairly constant emotions during the book, but fear and sadness change a bit. There's more fear earlier in the book, aligning with how Sancho is much more timid early on. Sadness is all more prevalent early on, aligning with Don Quixote and Sancho continuously getting injured. Sadness also near the ending of the book, where Don Quxiote must give up his knightly ways, and ultimately dies.