# Homework 2C - DATA-312

Jeffrey Williams

08, April 2022

**Abstract**

This writeup explores the

### ANALYSIS OF SENTIMENT IN DICKENS - USING THE BING LEXICON

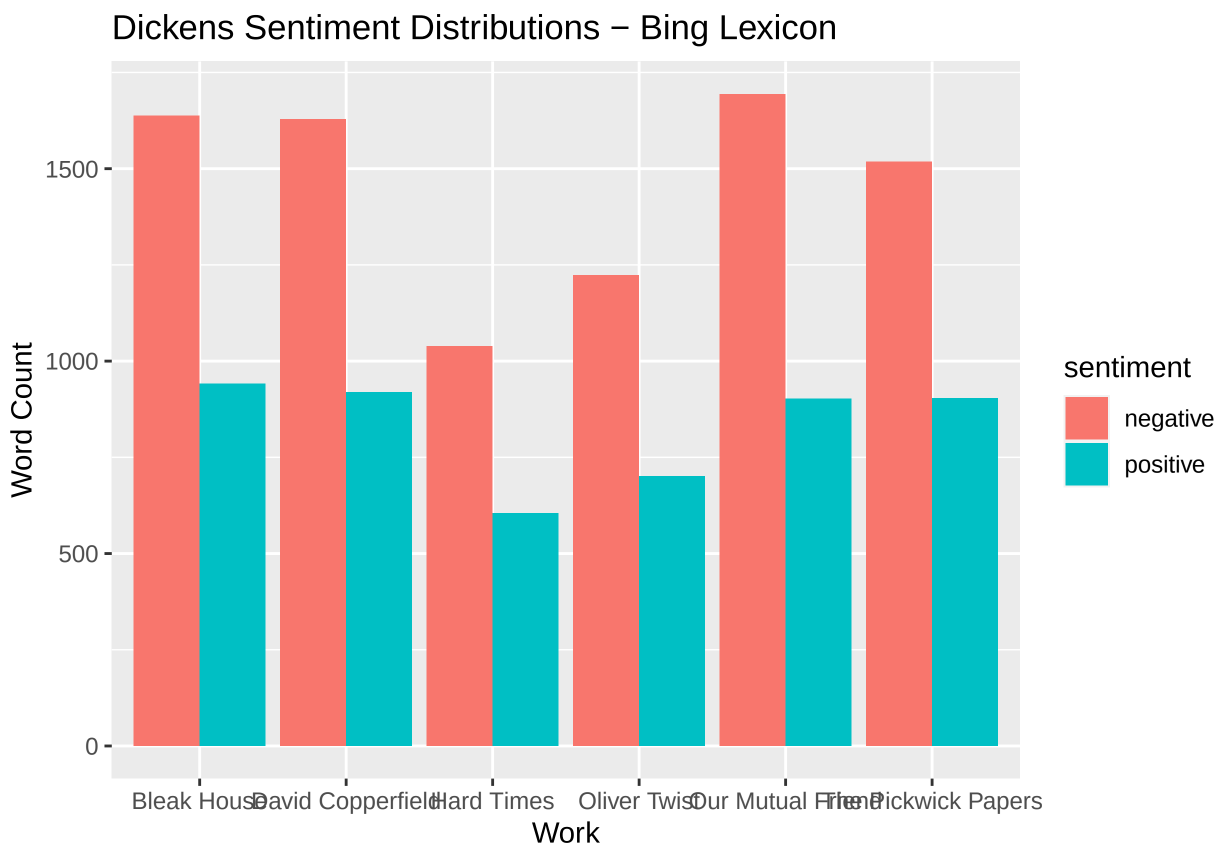
The Bing lexicon consists of 6786 words, sorted into two categories, negative and positive, based on their perceived connotations. Such a lexicon was applied to the dataframe of the select works of Dickens in an effort to begin to understand the balance of sentiment (or lack thereof) thematic in his work, both in terms of individual works and, if possible all of them. Is Dickens prone to writing generally negative works?

Bearing in mind the nature of Dickens as a serial novelist, which explains the similarities in proportion for most of the works evaluated, it is obvious that there is a signifant overbearing of negative sentiment over positive. For each individual work, the count of words aligning with a negative sentiment per the Bing lexicon are significantly higher than words that are classified as negative.

```r
tidy_dickens <- tidy_dickens %>% inner_join(bing_sentiments) %>% group_by(Name)
```

```
## Joining, by = "word"
```

```r
tidy_dickens %>%
  count(sentiment) %>% ggplot(aes(Name, n, fill=sentiment)) +
  geom_col(position='dodge')  +
  labs(title="Dickens Sentiment Distributions - Bing Lexicon",
       x = "Work",
       y = "Word Count")
```



It is strongly implied here the conclusion that sad themes are recurrent in the work of Dickens. However, this assertion could be even more strongly substantiated by a chi-square test.

```
# Contingency table
ct <- tidy_dickens %>% inner_join(bing_sentiments) %>% group_by(Name) %>%
  count(sentiment) %>% pivot_wider(names_from = sentiment, values_from = n) %>%
  column_to_rownames(var = "Name")
```

```
## Joining, by = c("word", "sentiment")
```

```
# Chi-sq test
chSq <- chisq.test(ct)
chSq
```

```
##
##  Pearson's Chi-squared test
##
## data:  ct
## X-squared = 4.0559, df = 5, p-value = 0.5414
```

```
chSq$observed
```

```
##                       negative positive
## Bleak House               1637      941
## David Copperfield         1628      919
## Hard Times                1038      605
## Oliver Twist              1223      701
## Our Mutual Friend         1694      902
## The Pickwick Papers       1518      903
```

```r
chSq$expected
```

```
##                     negative positive
## Bleak House         1643.195 934.8047
## David Copperfield   1623.436 923.5639
## Hard Times          1047.234 595.7658
## Oliver Twist        1226.341 697.6588
## Our Mutual Friend   1654.668 941.3317
## The Pickwick Papers 1543.125 877.8752
```

```r
chSq$stdres
```

```
##                       negative   positive
## Bleak House         -0.2816649  0.2816649
## David Copperfield    0.2084625 -0.2084625
## Hard Times          -0.5051040  0.5051040
## Oliver Twist        -0.1708915  0.1708915
## Our Mutual Friend    1.7834267 -1.7834267
## The Pickwick Papers -1.1705167  1.1705167
```

```r
afinn_sentiments <- get_sentiments("afinn")
```

The Chi-Squared test identifies a p-value of 0.541393, meaning that there is no level of significance here from a more numerical standpoint. With this lack of significance in mind, it can be concluded, therefore, that it is a typicality in Dickens to write generally negative pieces. What a sad individual he was!

**ANALYSIS OF SENTIMENT IN DICKENS - USING THE AFINN LEXICON**

Similarly to the Bing lexicon, the AFINN lexicon is used to evaluate the sentiment of a variety of words. In this case, the AFINN lexicon includes 2477 words from the English language. The key difference though is that rather than sorting individual words into different categories, AFINN instead assigns each included word an integer between -5 (most negative) and 5 (most positive). This is helpful in allowing us to understand the weight of a word's sentiment. In other words, in addition to showing that a word is negative or positive, it also helps us understand how negative or how positive a word is.

Here, we apply the AFINN lexicon to our dataframe consisting of