Decoding Hewins’ Books for the Young

Victoria Chang, Aaron Gong, and Shelley Kim

Project Advisor: Nicole Goodman

Introduction

- Caroline Hewins is considered one of the “founding mothers” of the field of children’s literature, defining the canon of the genre with her handbook *Books for the Young* (1882).
- The **goal of our research** is to use statistical methods, primarily logistic regression, to quantitatively examine the distinguishing features of children’s fiction literature compared to a corpus of fiction literature from the same time period.

Data Description

- Hewin’s corpus was provided by Dr. Rebekah Fitzsimmons and consisted of 1,089 books of varying genres. Of these 1,089 books, 329 were considered to be fiction, and 176 had transcribed versions available for use. These 176 books range from year 1808-1882.
- The comparison corpus of non-canon-books was derived from the University of Chicago Textual Optics Lab’s U.S. Novels corpus. We selected the earliest 176 books from this corpus, ranging from year 1880-1894, to compare against the Hewins corpus. This yields 352 books total for our analysis.

Methods

- **Word frequency analysis** is performed for each corpus as a whole, as well as for each book in each corpus.
  - Helped establish the “signature” of the Hewins corpus and the Chicago corpus, respectively.
  - Extracted key features of each book for the logistic regression model.
- **NRC sentiment analysis** is performed for each book in each corpus.
  - Utilized the NRC sentiment dictionary, which contains 10 different sentiments, 5 positive and 5 negative.
  - Computed the prevalence of each of the 10 sentiments in each book.
    - This is done by classifying the top 1000 most frequent words in each book into 10 sentiments weighted by word frequency, and computing the proportion of each sentiment.
- **Logistic regression model** with L2 regularization is performed using LightSIDE software.
  - Feature variables: selected metadata (author gender & nationality), selected word frequencies, proportions of word sentiments.

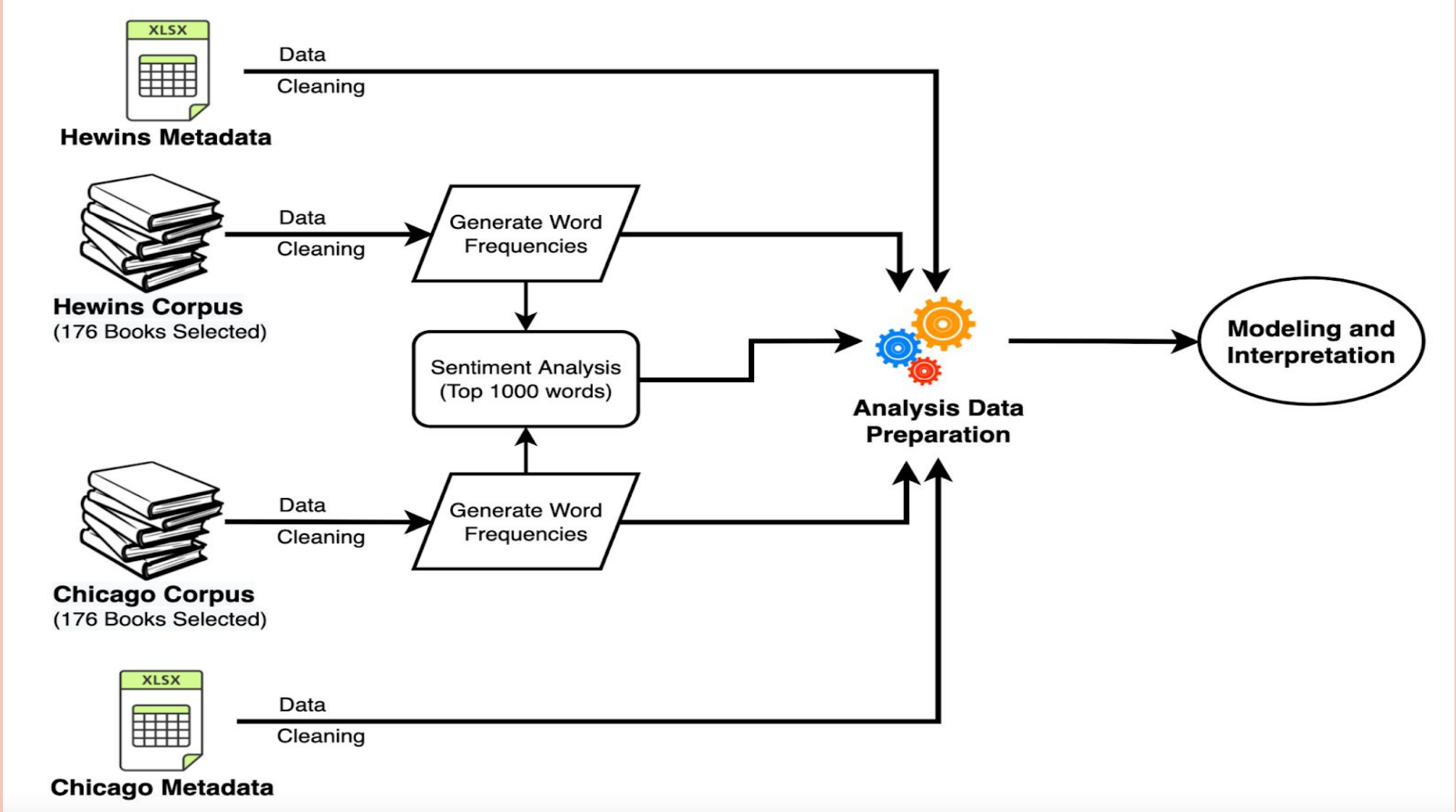


Figure 5: Flow Chart of Research Approach

Modeling

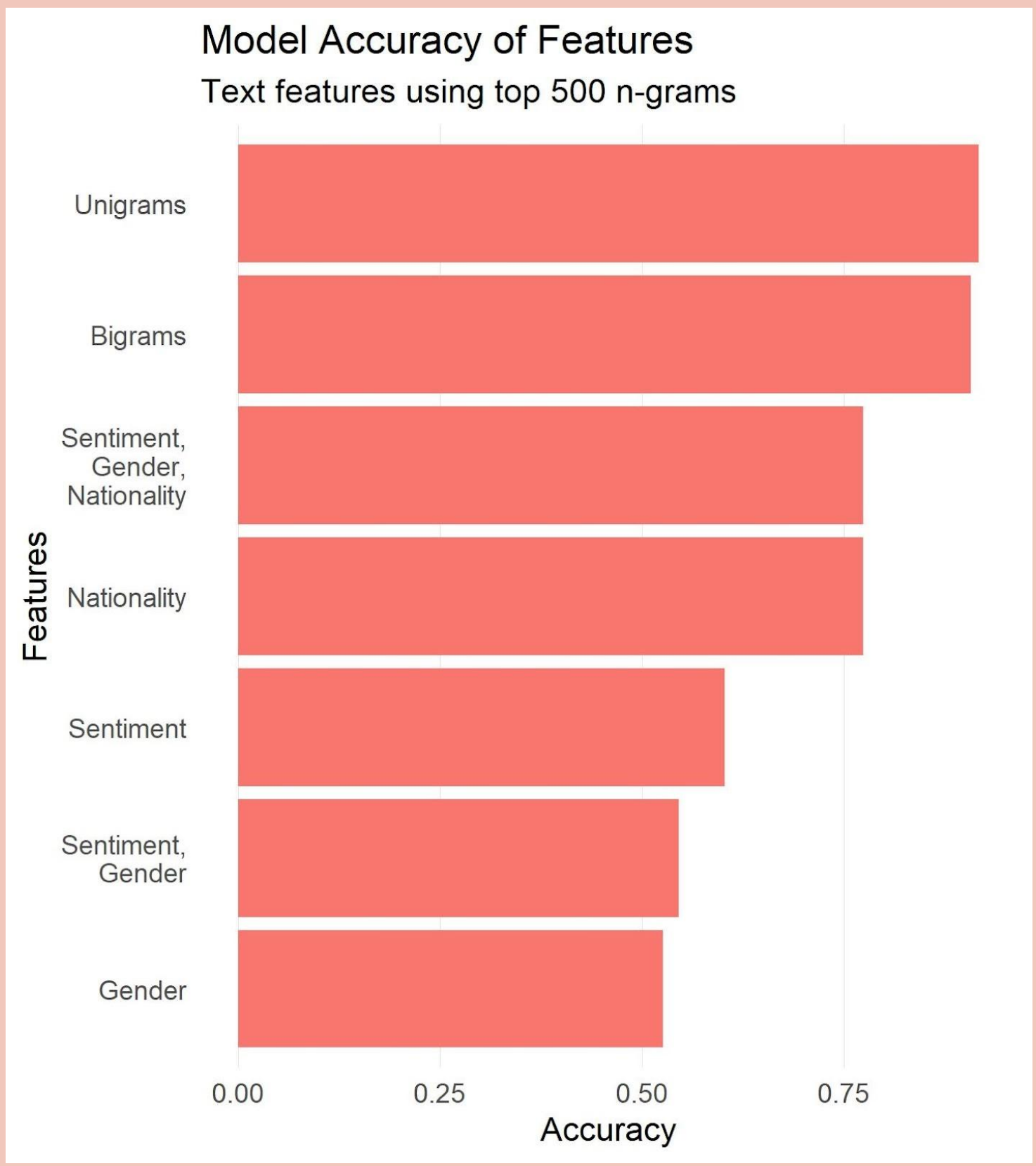


Figure 7: Comparison of different models by accuracy

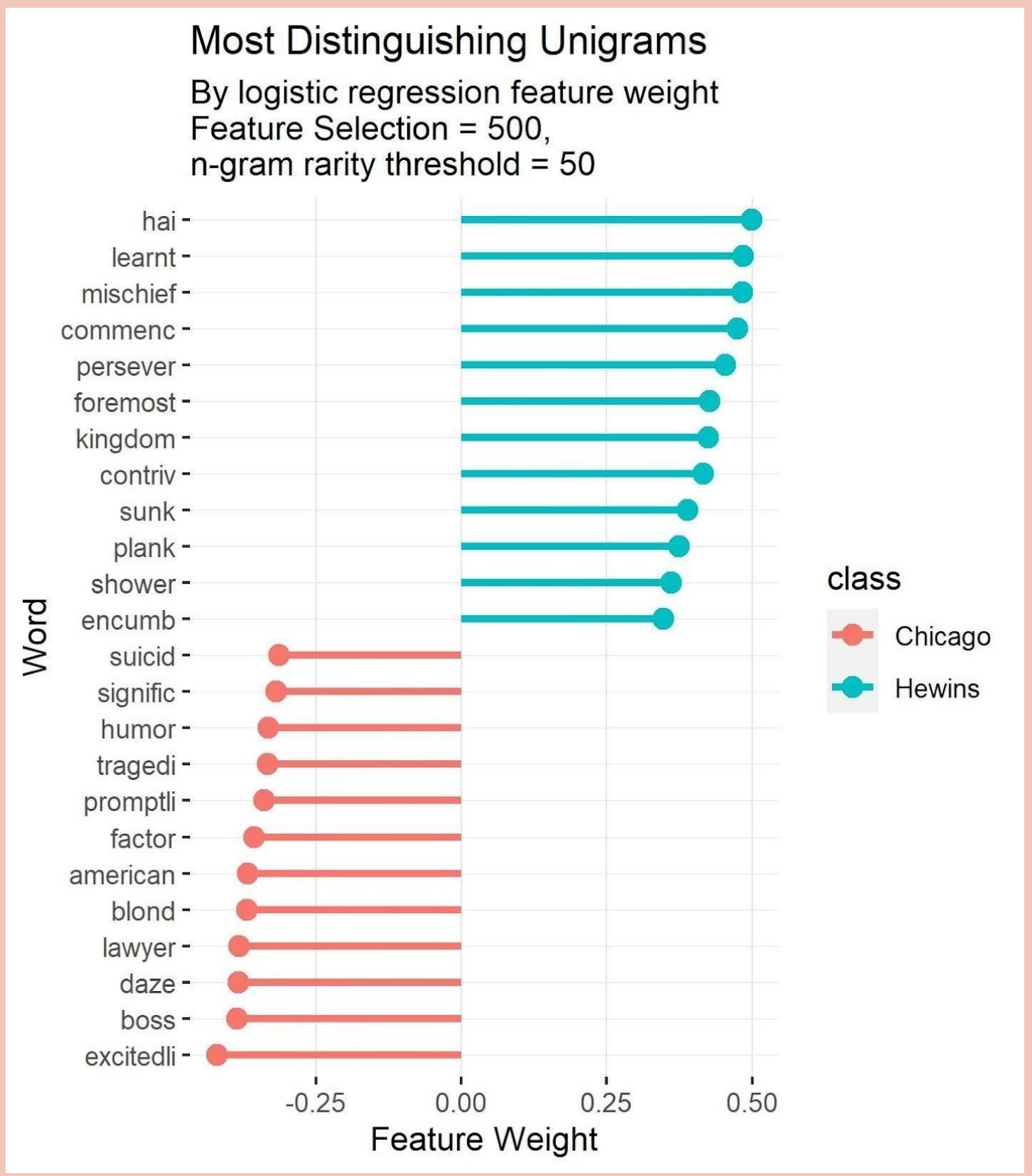


Figure 8: Most distinguishing features of the unigram logistic regression model.

- The n-gram models were the highest performing, with unigrams being the most accurate. The unigram model used in Figure 8 selected the top 500 unigrams and yielded 91.7% accuracy.
- Bigrams features of non-selected novels had words concerning romance and marriage, which were absent from the selected novels corpus.
- Author nationality was the best-performing meta-feature with 77.4% accuracy.
- Words associated with Hewins’ books are more adventurous and playful than those of the comparison corpus, while words such as “suicide”, “tragedy”, and “lawyer” are highly-weighted for the comparison corpus.

Exploratory Data Analysis

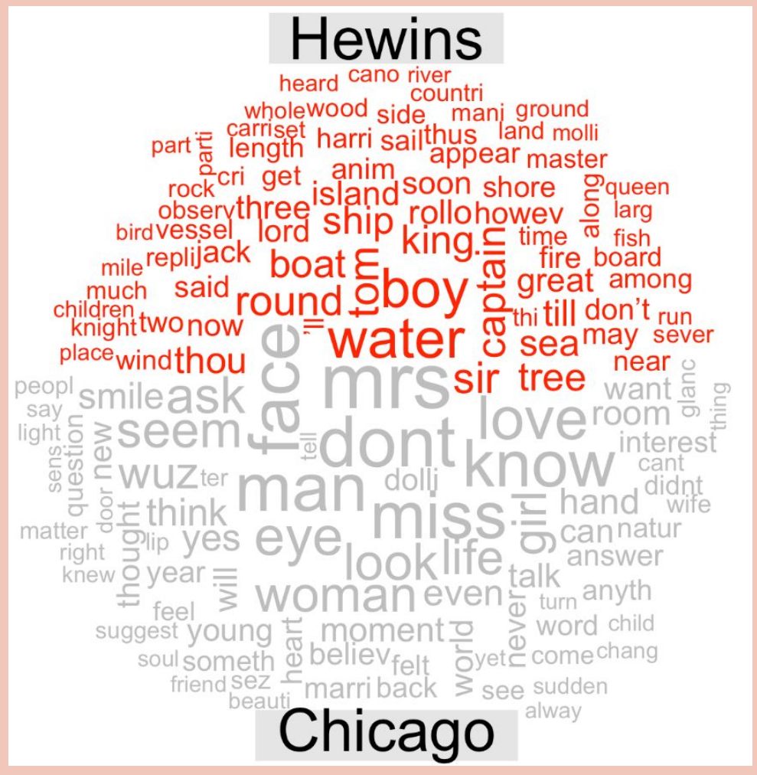
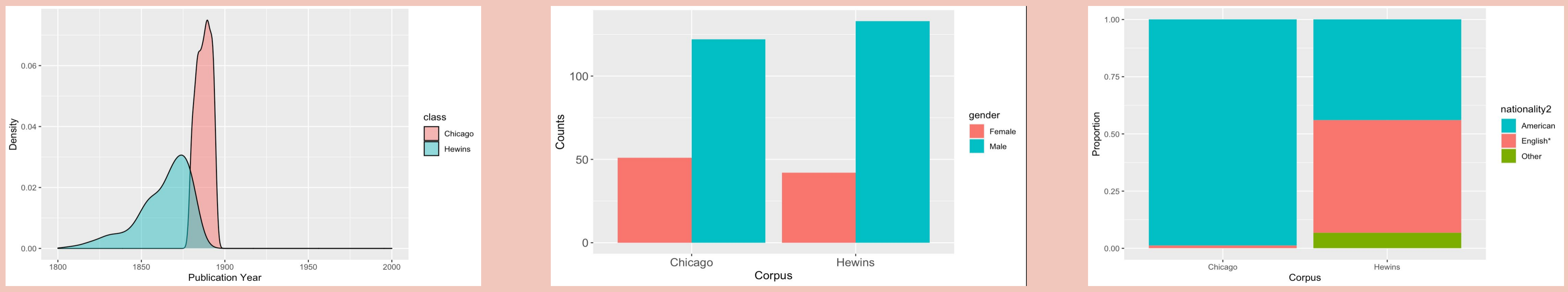


Figure 1: Comparison word cloud
Comparison cloud for most frequent words from each corpus

- Books in the Hewins corpus contain more adventurous and story related words such as king, queen, boat, ship, captain, island and sea.
- “boy” is one of the top Hewins words, so it is likely Hewins selected more books with male characters.
- Books in the Chicago corpus address “man”, “girl”, “woman”, “mrs” more frequent than Hewins corpus.



Figures 2,3, & 4: Plots of author metadata by corpus
Left: Density plot of publication year; **Middle:** Bar plot of author gender; **Right:** Bar plot of author nationality

- English*: English, British & Scottish
- Other: French, German & Canadian

Sentiment

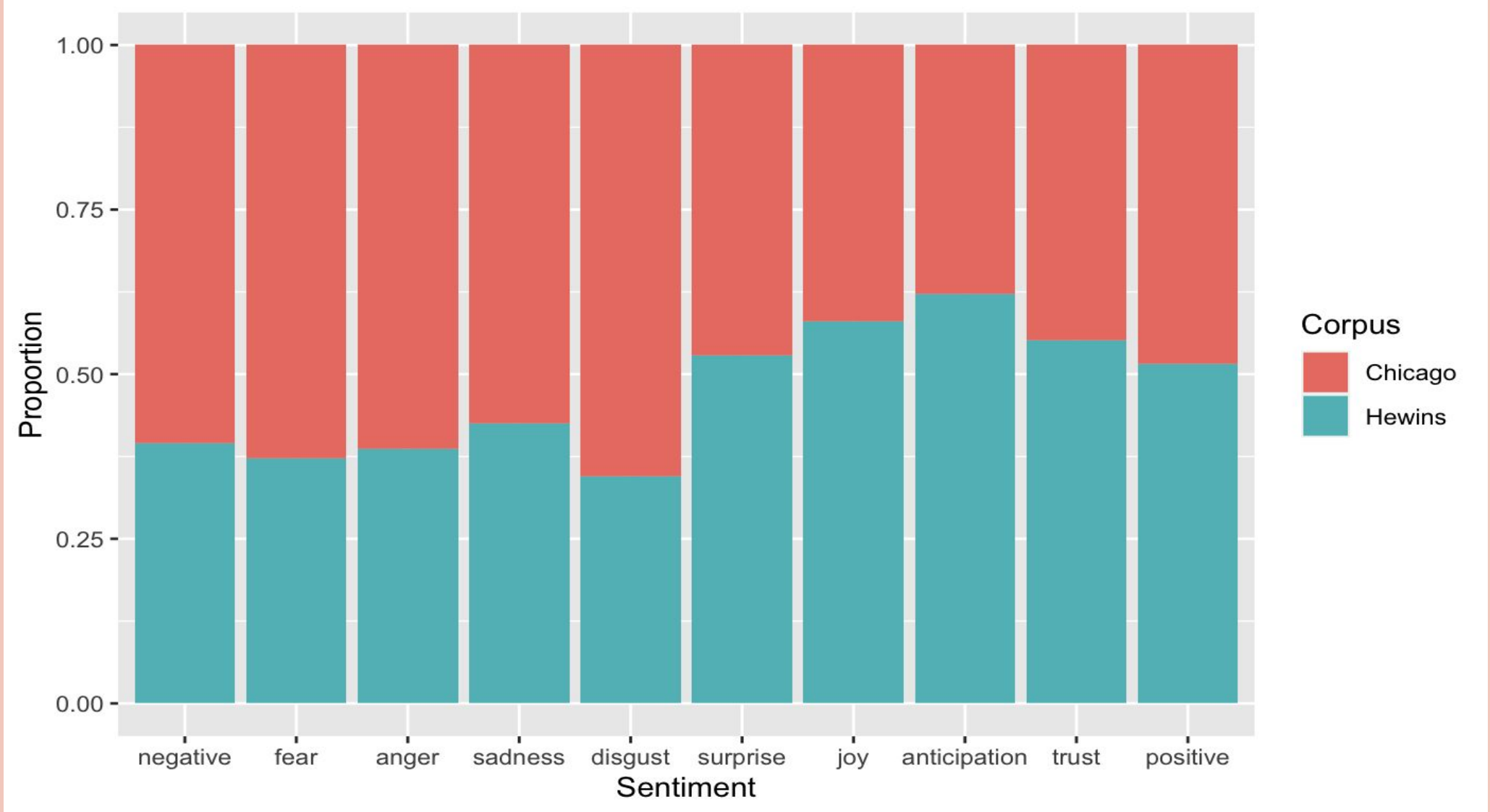


Figure 6: Proportional bar plot showing prevalence of sentiments by corpus

Based on the top 1,000 words in each corpus:

- The Hewins corpus is more likely to contain words with positive sentiments (e.g. surprise, joy, anticipation, and trust).
- The Chicago corpus is more likely to contain words with negative sentiments (e.g. fear, anger, sadness, and disgust).

Conclusion

Our research contributes to the groundwork for future explorations into Hewins’ influence on the field of American children's literature in the following ways:

1. We extracted fiction books in the Hewins corpus and compared them to fiction books from the same time period.
2. We identified differences in sentiment between the Hewins and Chicago books.
3. We obtained useful features which provided insight into how language differs between the two corpora.

References

- The University of Chicago Textual Optics Lab. *U.S. Novel Corpus*, 2021.
- Underwood, Ted. *Distant Horizons*. The University Of Chicago Press, 2019.