Was it always “Happily Ever After”?

A Cross Corpora Analysis between Children’s Books Lists of Caroline Hewins and Anne Carroll Moore

By Elliot Buera, Eric Shau, Michael Zheng, Patrick Zhu
Research Advisor: Aaditya Ramdas

Introduction

- Caroline Hewins and Anne Carroll Moore are renowned figures in children's literature
- Hewins' “Books for the Young” (1882) and Moore's “A List of Books Recommended for a Children's Library” (1902) were vital in standardizing children’s literature and shaping conceptions of children's reading materials
- This research explores how to define and analyze whether children's’ books have happy endings
- **Key Research Question:** Can sentiment analysis of book endings reveal quantitative differences between Hewins’ and Moore’s children book lists?

Data & Data Cleaning

- Data consists of 2 corpora of Hathitrust OCR text files and metadata of said texts
- Metadata includes the title, author, publication date, category, etc.
- We did initial data cleaning (inspired by Prof. David Brown @ CMU) and manually deleted irrelevant text in beginnings and endings (e.g., Table of Contents, Footnotes, etc.)

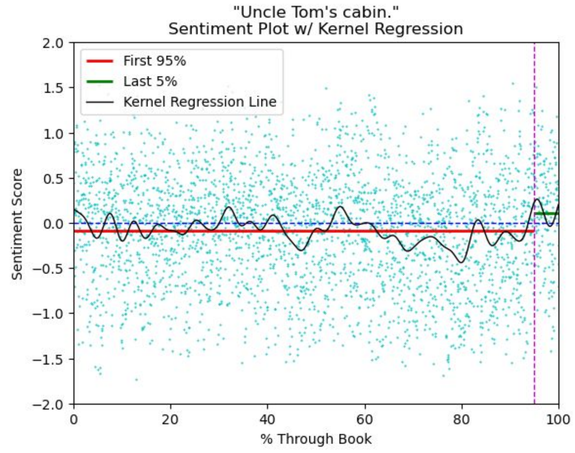
Methods

- Re-categorized corpora to standardize them
- Used tabularisai’s robust-sentiment-analysis model on chunks of each text [1, 2]
- Identify the last 5% of the book as the ending of the book [3]
- Converting the sentiment difference between the mean sentiment of last 5% and first 95% to a modified “z-score” (formula below)
- Define positive ending as z-score > 0.125, negative ending as z-score < -0.125, neutral ending as -0.125 ≤ z-score ≤ 0.125

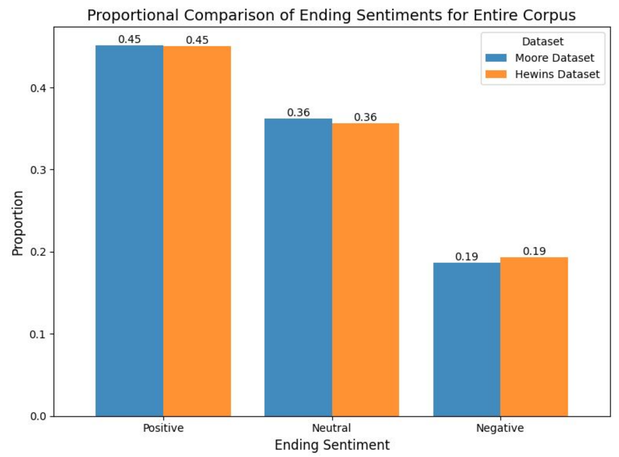
$$\text{modified z-score} = \frac{\overline{\text{last } 5\%} - \overline{\text{first } 95\%}}{\text{combined SD}}$$

$$\text{combined SD} = \sqrt{\frac{(n_{\text{first}} - 1)\text{SD}_{\text{first}}^2 + (n_{\text{last}} - 1)\text{SD}_{\text{last}}^2}{n_{\text{first}} + n_{\text{last}} - 2}}$$

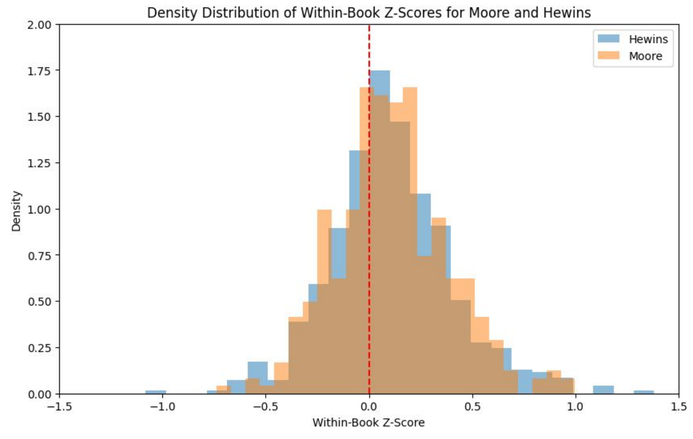
Results



- Example of book with “happy” ending
- Sentiment score normalized across the length of the book
- An RBF kernel regression with gamma=0.1 is also plotted



- Proportions of ending sentiments between each corpus are almost the exact same!



	Hewins	Moore
Mean	0.110	0.111
SD	0.299	0.278
Median	0.089	0.099

- Distribution and statistics of the within-Book Z-score of Moore and Hewins’ lists for the selected categories
- Very similar distributions between corpora (approximately normal too)

Conclusion

- Distributions of score differences and happy endings are surprisingly similar
- Ending type proportions vary in comparing categories
  - Generally, positive and neutral ending proportions outweigh negative ending proportions in both corpora

References

[1] “tabularisai/robust-sentiment-analysis · Hugging Face,” *Huggingface.co*, 2024. https://huggingface.co/tabularisai/robust-sentiment-analysis

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” *arXiv.org*, Feb. 29, 2020. https://arxiv.org/abs/1910.01108v4

[3] A. Zehe, M. Becker, L. Hettinger, A. Hotho, I. Reger, and F. Jannidis, “Prediction of Happy Endings in German Novels based on Sentiment Information,” University of Würzburg, 2016. Available: https://ceur-ws.org/Vol-1646/paper2.pdf