

Do Wars affect a Story's Emotional Arc? Books as a Collective Coping Mechanism

Erika Gutierrez, Eric Frey, Davis Thomas

February 2023

1 Introduction

It is a common anecdote that books can serve as a source of refuge. Furthermore, there is ample evidence suggesting the benefit of reading fiction in aiding stress relief and even as a coping mechanism when needing an escape from difficult situations. As we will see in this paper, fiction stories can take many forms. However, some of the best works of fiction are the ones where readers can relate and connect deeply with the emotional experience produced by the plot of the book. These ideas lead us to want to understand if there is a relationship between emotional shocks to society and common emotional arcs found in the books published around the same time as those shocks. To test this, we use World War I (WWI) as an example emotional shock to society and compare the frequency of the emotional arcs shapes in the books published leading up to the war and after it.

2 Data

2.1 GoodReads

In order to obtain a list of books that were written before and after WWI, we webscraped from Goodreads. GoodReads is a website that allows users to catalog books, rate books, write book reviews, and receive recommendations about what books to read next. Within GoodReads there are webpages called shelves, where if users tag a book with a particular label, it will appear on that shelf. We utilize decades shelves, which contain books tagged as a specific decade. For the 1890s, 1900s, 1910s, 1920s, and 1930s shelves, we scrape the first ten pages of each shelf, as the books within a shelf are sorted in decreasing order by how often a book gets tagged as that label. For each book, we scrape book title, author, and year the book was first published. This "year published" field is the most important, as the Gutendex (the source of our book text data) does not contain data on when the book was originally published (otherwise, we would not have to rely on GoodReads and could simply use the Gutendex). Once we have scraped each shelf, we filter for books that were actually published in the relevant decade as some users tagged books where the *setting* was in the 1910s as being a 1910s book, rather than when it was written. After scraping GoodReads shelves, we end up with metadata for approximately 1,200 books written between 1890 and 1939.

2.2 Project Gutenberg

Project Gutenberg is a free online library of over 60,000 electronic books. The project was started by Michael S. Hart, the inventor of the e-book, in 1971. Today the project is run by a system

of volunteers who continually select and digitize books available in the public domain. For our project we access the books on Project Gutenberg through Gutendex, a JSON web API for Project Gutenberg ebook metadata. Using the list of books written before and after WWI mentioned in the prior section, we queried the Gutendex API to retrieve the relevant text files that would become the text corpus used in our analysis. In our query we specified that the books had to belong to the fiction genre, written in English, and match the author’s name and book title. In the end we were only able to download 193 text files. And then after filtering for books between word counts of 10,000 and 100,000 words to ensure books were neither too complex nor not complex enough, we obtained 139 books written from 1890 to 1938. The fact that we could not download the full 1,2000 books from our original list may be due to the books not being in the public domain. Project Gutenberg only publishes books where the U.S. copyright has expired, which can be typically 70 years after the author’s death. Moreover, we obtained a much greater proportion of books written before WWI than after WWI.

3 Methodology

3.1 Sentiment Analysis to Retrieve Emotional Arcs

An emotional arc is the trajectory of the emotional journey that a story takes its audience through. It is a crucial component of storytelling, as it can significantly impact the way the audience perceives and remembers the story.

From our corpus of books, we filter for books that have a minimum length, to filter out short stories. We also pre-process the data by removing certain headers and footers that Project Gutenberg adds to the data, and extract the main body of text.

To reconstruct the emotional arc from the data available to us, we analyze the sentiment of ‘windows’ of text, consisting of 10,000 words, which is then moved through the entire story text, and a predetermined number of observations is taken. The number of observations taken and the temporal resolution of our time series data are both important for our objective, and since we need to compare the time series across different books, we have an additional requirement of consistency across books. So we selected a fixed window size k , and we vary the amount that each window has to move, and we calculate this using the following equation from [1]. For a time series of length l and a book of N words, the overlap of the windows is

$$(N - k - 1)/l \tag{1}$$

We calculate the sentiment of each window using the LabMT dataset, which has a large dictionary of words scored through surveys and also includes a variance score for the value rated for each word in dictionary. The collected observations are then aggregated in a matrix, where each row represents the emotional arc of a single story in our corpus.

In the figure below, we show an example of what a story looks like after calculating the sentiment throughout the story. Note that the story has a clear downward trend after the initial rise, only to pick back up at the conclusion. To demonstrate the sentiment variance within the story, we provide an excerpt from the peak point of the story at around the 15% mark:

They must have been dining, sir, and seemed more inclined to lark about than to listen to good music. The moment they entered the box, they came out again and called the box-keeper, who asked them what they wanted. They said, ‘Look in the box: there’s no one there, is there?’ ‘No,’ said the woman. ‘Well,’ said they, ‘when we went in, we heard a voice saying THAT THE BOX WAS TAKEN!’”



Figure 1: Emotional story arc of Gaston Leroux’s Phantom of The Opera

M. Moncharmin could not help smiling as he looked at M. Richard; but M. Richard did not smile. He himself had done too much in that way in his time not to recognize, in the inspector’s story, all the marks of one of those practical jokes which begin by amusing and end by enraging the victims. The inspector, to curry favor with M. Moncharmin, who was smiling, thought it best to give a smile too. A most unfortunate smile! M. Richard glared at his subordinate, who thenceforth made it his business to display a face of utter consternation.” [2]

Note how words like ”lark”, ”good”, ”music”, and ”smile” likely contribute to the peak sentiment of the story. Compared to the low-point of the story towards the end, we can clearly see a sentiment difference:

I have prayed over his mortal remains, that God might show him mercy notwithstanding his crimes. Yes, I am sure, quite sure that I prayed beside his body, the other day, when they took it from the spot where they were burying the phonographic records. It was his skeleton. I did not recognize it by the ugliness of the head, for all men are ugly when they have been dead as long as that, but by the plain gold ring which he wore and which Christine Daae had certainly slipped on his finger, when she came to bury him in accordance with her promise.

The skeleton was lying near the little well, in the place where the Angel of Music first held Christine Daae fainting in his trembling arms, on the night when he carried her down to the cellars of the opera-house.

And, now, what do they mean to do with that skeleton? Surely they will not bury it in the common grave! ... I say that the place of the skeleton of the Opera ghost is in the archives of the National Academy of Music. It is no ordinary skeleton.

Here words like ”mortal”, ”grave”, ”crime”, and ”skeleton” contribute to the relatively low sentiment score.

The data in this form can be difficult to analyse and draw meaningful inferences from, as each emotional arc has significant variance in the value, and a lot of noise that can be difficult to isolate. A human observer who is familiar with the plot of the book will be able to identify inflections in the

arc, and relate them to certain events in the book, but we need to isolate some commonality from our corpus which we can then use for our analysis.

3.2 Singular Value Decomposition to Retrieve Modes of Emotional Arcs

This commonality can be obtained from our data using Singular Value Decomposition (SVD), which is a form of Principal Component Analysis (PCA).

PCA is a statistical technique used to reduce the dimensionality of a dataset by finding a set of new variables, called principal components, that capture the most significant information in the original dataset.

It works by identifying the linear combinations of the original variables that have the highest variance, which are the directions in the data that capture the most information. These linear combinations are the principal components, and we use these to represent our data using less dimensions than our original dataset.

SVD is used to calculate the principal components by decomposing the data matrix A into a linear combination of three matrices: U , Σ , and V . [3]

The U matrix is an orthogonal matrix of size $m \times m$, where m is the number of rows of A . The columns of U are the left singular vectors of A , which form an orthonormal basis for the column space of A . Each left singular vector represents a direction in which the data in A varies, and these are ordered from highest to lowest variance. In this sense, U can be thought of as a transformation matrix that maps the columns of A onto a new set of basis vectors.

The Σ matrix is a diagonal matrix of size $m \times n$, where m and n are the number of rows and columns of A , respectively. The diagonal entries of Σ are the singular values of A , which capture the importance of each left and right singular vector. The first singular value represents the importance of the first left and right singular vectors in capturing the variation in A , the second singular value represents the importance of the second left and right singular vectors, and so on. In this sense, Σ can be thought of as a scaling matrix that adjusts the magnitude of each left and right singular vector.

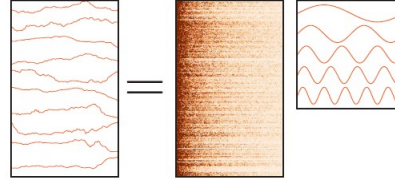
The V matrix is an orthogonal matrix of size $n \times n$, where n is the number of columns of A . The columns of V are the right singular vectors of A , which form an orthonormal basis for the row space of A . Each right singular vector represents a pattern in the data in A , with the first vector capturing the most dominant pattern, the second capturing the second-most dominant pattern, and so on. In this sense, V can be thought of as a transformation matrix that maps the rows of A onto a new set of orthonormal basis vectors.

$$A = U\Sigma V^T = WV^T \quad (2)$$

In the context of the emotional arc matrix A , each row of A represents the time series of an individual book, and each column the percentage point of the book at which that sentiment observation was recorded. Different interpretations are possible for the three different matrices that are created after decomposition, however here we shall consider the columns of the matrix V as the basis vectors for the sentiment time series in the rows of A . This will now be referred to as the modes.

Looking at the values inside the ordered matrix Σ , the diagonal entries are the singular values of the original matrix A . The singular values represent the scaling factors applied to the modes during the matrix transformation. Larger singular values indicate a more important contribution of that mode to the matrix A . The Σ matrix allows us to measure the relative importance of each singular vector in the transformation of the data. The first few principal components, corresponding to the largest singular values, capture most of the variation in the data, allowing for dimensionality

reduction and visualization of the data. We will combine the U and Σ into the coefficient matrix W , which now represents the mode coefficients.



$$A = W V^T$$

Figure 2: Visualization of the Modes and Mode coefficients after SVD

3.3 Clustering

We then use K-means clustering on the normalized mode coefficients to cluster the different books based on the emotional arc that best describes the book. We select the number of clusters to be 5, as that yielded the best visual results. We also tested an agglomerative clustering algorithm but K-means with normalized weights gave us the clearest groupings of stories. Figure 3 plots the scaled sentiments of the books by cluster. We can see that there is a common emotional arc within each story grouping as indicated by the trendline.

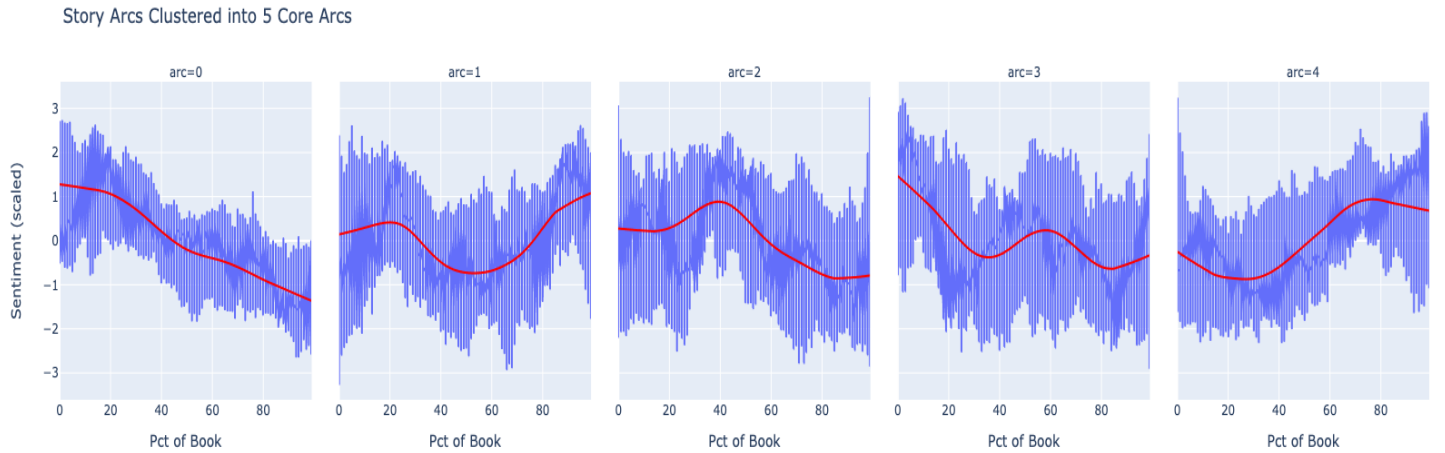


Figure 3: The 5 resulting emotional story arcs after clustering the coefficients of the stories

3.4 Hypothesis Test

To understand the relationship between the shock of WWI and the frequency of the emotional arcs we make use of the chi-square hypothesis test. This test is commonly used when determining if there is a significant association between categorical variables - in this case the emotional arcs of books. Our null hypothesis is that there was no change between the frequency of the emotional arcs before and after the war.

We group books by publishing date: books before 1914 are classified as "Before WWI", books from 1914-1918 are classified as "During WWI", and books from 1919-1938 are classified as "After WWI". After omitting the "During" category, we are left with a sample size of 127 books which we use to examine the differences between the proportion in story arcs.

4 Results

Below we categorize the uncovered emotional arcs and state the change in their frequency before and after WWI:

- Arc 0: Tragedy (fall) - more frequency post WWI
- Arc 1: Cinderella (rise fall rise) - less frequency post WWI
- Arc 2: Icarus (rise fall) - more frequency post WWI
- Arc 3: Oedipus (fall rise fall) - More frequency post WWI
- Arc 4: Rags to Riches (rise) - less frequency post WWI



Figure 4: Proportion of stories that belong to an arc (relative to the Before/After WWI group)

Based on the results of the chi-squared test, we can not state with certainty that WWI had an impact on the frequency of books associated with the uncovered emotional arcs. The p-value of the test was .804 which is greater than our desired significance level of .05.

Table 1: Chi-Squared Test for statistical significance of WWI on emotional story-arcs

	Observed					Total
	Arc 0	Arc 1	Arc 2	Arc 3	Arc 4	
Before WWI	20	17	20	18	18	93
After WWI	8	4	8	9	5	34
Total	28	21	28	27	23	127

	Expected					Total
	Arc 0	Arc 1	Arc 2	Arc 3	Arc 4	
Before WWI	20.50	15.38	20.50	19.77	16.84	92.99
After WWI	7.50	5.62	7.50	7.23	6.16	34.01
Total	28	21	28	27	23	127

Degrees of freedom = 4
Chi-squared statistic = 1.622
P-value = .804

Due to time and data constraints our sample size was very small, consisting of only 127 books. We believe this had a significant affect on the results, as one of the assumptions of the Chi-squared test is that each category must be sufficiently large, which our data does not satisfy. If more of the books that we scraped from Goodreads were available on Project Gutenberg perhaps we could identify a clearer relationship. However, we also note that there are many other influences that come into play when an author publishes a book such as the region that the author grew up in or lived in at the time of writing the novel, the degree to which they were affected by the war, personal motivations that pushed them to writing the story in question, events in the author’s life, to name a few.

5 Conclusion

Upon reflection, we identified several improvements that could be been made to make our analysis better. For example, making sure the books we selected were sufficiently randomized. To create our book list we choose the top ten pages of search results on GoodReads, which could be prone to selection bias if it only shows the most popular books. Additionally, we are assuming that all books are independent of each other in order to apply the chi-squared test when this might not be the case. Lastly, our sample size is too small, we should find a way to include more books into the analysis. A way to improve sample size if more time was available would be to first get a list of available books on Gutendex within a time-frame of when the author was alive (as this is the closest data to publishing date), look-up the books’ publishing dates on Goodreads, filter for books published around WWI, and then download the books from Gutendex. It appears that the Gutendex database is less complete than Goodreads, so starting from what Gutendex has, rather than from what Goodreads has, may yield a larger number of books. In this project, we focused solely on the change in emotional arc, but there are many other ways one could examine the differences in literature that could yield interesting results. For example, perhaps the topics discussed within a book or how the main character is presented could also change throughout time.

Through the use of open source data and natural language processing techniques, we were able to extract valid emotional arcs of books. Having these tools at our disposal allows us to ponder

questions like the one explored in this paper. Although our results were inconclusive, our analysis provides a foundation towards contributing to the understanding of how collective traumas can shape cultural production and the role of literature as a coping mechanism for individuals during difficult times.

References

- [1] A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, and P. S. Dodds, “The emotional arcs of stories are dominated by six basic shapes,” *CoRR*, vol. abs/1606.07772, 2016.
- [2] G. Leroux, *Phantom of The Opera*. Penguin, 1909.
- [3] G. Gundersen, “Singular Value Decomposition as Simply as Possible.” <https://gregorygundersen.com/blog/2018/12/10/svd/>, 2018. [Online; accessed 02-03-2023].