Stream of Consciousness Literature: A 'Joyceless' Linguistic Landscape*

Exploring Word Frequency, Sentiment Value and Mental Health Themes in the Works of Joyce, Woolf, Proust, Mansfield and Eliot from Project Gutenberg

Quang Mai

April 24, 2024

This project focuses on understanding the language used by renowned stream of consciousness (SOC) authors James Joyce, Virginia Woolf, Marcel Proust, Katherine Mansfield and T.S Eliot. By conducting word frequency analysis and sentiment analysis on these authors' nine novels, we aim to uncover shared linguistic patterns and gain insights into the authors' mental states. Our analysis of the SOC genre shows that higher frequency words are mostly negative and carry sadness or fear-inclined connotations, with clear linguistic differences between female and male authors. With these trends uncovered, we attempt to offer insights into themes of self-identity, anxiety, disassociation and existential contemplation within Western society and its literary circle from late 19th to mid-20th century.

Table of contents

*Code and data are available at: https://github.com/ponolite/stream-consciousness-language.git

1 Introduction

Stream of consciousness (SOC) is a narrative technique that aims to capture the continuous flow of thoughts, feelings, and sensations experienced by a character without conventional organization or punctuation (Bernini and Fernyhough 2022). It mirrors the unpredictable and interconnected nature of human thought processes, often revealing the emotional well-being and psychological struggles of the character's mind in an intimate and unfiltered manner (Long and So 2016) (Nyongesa 2023). As such, this paper has mined the texts of a total of nine novels from the volunteer archive, Project Gutenberg, to examine the mental health themes of famous stream of consciousness authors, namely by Joyce, Woolf, Proust, Mansfield and Eliot, from the modernist era of literature, spanning from late 19th century to mid-20th century ("Project Gutenberg," n.d.).

By analyzing these textual datasets through word frequency and sentiment analysis, this paper seeks to pose and answer crucial questions: *What are some important factors contributing to this relationship between mental health, disassociation and stream of consciousness? Moreover,*

how does this relationship vary differently across differnet demographics of authors, for instance, authors with different genders? Understanding these dynamics is crucial in having an informed understanding of the West's late 19th to mid-20th century literature and even socio-political landscape, especially in regards to how authors and creative writers navigate and deal with then-taboo topics such as existential angst, mental health issues and anxiety

Thus, the estimand is the correlation between mental health-related themes in SOC literature, their appearance frequency and sentiment value. This is considered in terms of nine selected SOC texts only, namely Joyce's *A Portrait of the Artist as a Young Man and Chamber Music*; Woolf's *Mrs Dalloway and Jacob's Room*; Proust's *Swann Way*; Mansfield's *Bliss and The Garden Party*; and Eliot's *The Waste Land and The Love Song of J. Alfred Prufrock*. Through our analysis, we found that words that appear with high frequency are mostly negative in select SOC literature (-1.875 median on AFINN index). Sentiment-wise, analysis reveals a predominant 27% to 37% of words in SOC texts being negative and sadness-inclined (NRC index), and over 60% of words in these texts are negative (Bing index). Abstract and temporal nouns like "time," "day," "life," "moment," and "night" are most frequent, respectively appearing 870, 597, 567, 502 and 376 times in all texts combined. Gender-wise, male SOC authors are more likely to have negative words compared to female authors. These statistics correlate with SOC's non-linear nature, the genre's exploration of the unconscious mind (which often correlates to mental health themes), and its reflection of Western society in late 19th to mid-20th century.

To further understand the correlation between stream of consciousness texts and mental health themes, in Introduction, the paper briefly discusses the nature of stream of consciousness literature, relevant authors and the works that we have chosen to analyze. Subsequently, in Data and Results, we talk about the nature of the data obtained and analyze the results garnered from the data with suitable tables and charts. Next, Discussion provides further insights and future areas of study. Finally, Conclusion summarizes our main findings.

2 Data

The SOC texts used for analysis were sourced from Project Gutenberg under the library `gutenbergr` (Johnston and Robinson 2023) ("Project Gutenberg," n.d.). Data was generated, extracted and cleaned using the open-source statistical programming language R (R Core Team 2022), leveraging functions from `tidyverse` (Wickham et al. 2019), `tidytext` (Julia Silge and Robinson 2016), `rmarkdown` (Allaire et al. 2024), `dplyr` (Wickham et al. 2022), `ggplot2` (Wickham 2016), `scales` (Wickham, Pedersen, and Seidel 2023), `here` (Müller 2020a), `igraph` (J. Silge and Robinson 2006), `widyr` (J. Silge and Robinson 2022), `ggraph` (Pedersen 2024), `textdata` (Hvitfeldt 2022), `tm` (Feinerer, Hornik, and Meyer 2008), `here` (Müller 2020b), `kableExtra` (Zhu 2021), `arrow` (Richardson et al. 2024), and `knitr` (Xie 2014). More information on the dataset or the HTML textual data sourced from Project

Gutenberg is available in the documented datasheet available here within this folder, paper/datasheet_project_gutenberg.pdf, located within the paper's GitHub repository.

2.1 Measurement

Two central variables in this paper are:

- **Word Frequency:** This variable captures the repetition of a single word (unigram) or a pair of words combination (bigram) throughout a SOC novel text, providing us with a thematic understanding of SOC literature.
- **Sentiment Value:** This variable enables us to analyze how every word is usually perceived emotionally, whether it be a qualitative feeling or if it is conveyed through a numerical value.

Out of two variables used, the first one, 'Word Frequency' usually captured as `n/sum(n)` or `frequency` in datasets, is directly quantified through tokenizing the novel texts using the package `tidytext` and its function `unnest_tokens()` (Julia Silge and Robinson 2016). To do this, we downloaded all nine SOC texts from "Project Gutenberg" (n.d.), using functions such as `unnest_tokens()` from Julia Silge and Robinson (2016) to mine the texts, or separating them into individual words. Finally, we used `count()` to quantify the word count (`n`), and later divide the each word's count over the total word count of each author to garner the word frequency percentages (`n/sum(n)`).

The second variable used, 'Sentiment Value', is based on three English-based, general-purpose "word-emotion and word-polarity association lexicons", sourced from Mohammad and Turney's expansive research along with efforts from Finn Årup Nielsen and Bing Liu and collaborators (Julia Silge and Robinson 2016) (Mohammad and Turney 2013). The three general-purpose lexicon that contributes to this paper's sentiment analysis are (Julia Silge and Robinson 2016):

- 'AFINN' from Finn Årup Nielsen, which assigns each word a numerical value from '-5 to 5', judging its emotionality from negative to postive
- 'bing' from Bing Liu and collaborators, which assigns if a word is either 'positive' or negative'
- 'nrc' from Saif Mohammad and Peter Turney, which assigns a core emotional value to a word, such as 'fear', 'anger', 'sadness' or 'trust'

In terms of measuring 'Sentiment Value', all three general-purpose lexicons are compiled through crowd-sourcing and directly surveying the public on how each word is emotionally perceived. A survey sample of how the word 'startle' is compiled within the 'nrc' lexicon is presented below (Mohammad and Turney 2013):

(1) Which word is closest in meaning (most related) to startle?

- automobile

- shake
- honesty
- entertain

(2) How positive (good, praising) is the word startle?:

- startle is not positive
- startle is weakly positive
- startle is moderately positive
- startle is strongly positive

(3) How negative (bad, criticizing) is the word startle?

- startle is not negative
- startle is weakly negative
- startle is moderately negative
- startle is strongly negative

After the survey results are garnered, researchers average the answers to sort each surveyed word into pre-defined categories, specifically '-5 to 5' for 'AFINN', 'postive' or 'negative' for 'bing', and 'anger' or 'fear' for 'nrc'. With the continuous work of compiling these lexicons spanning years and decades (Mohammad and Turney 2013) comes these functions: `get_sentiments("bing")` and `get_sentiments("nrc")` and `get_sentiments("afinn")`. We can then use `inner_join` with these lexicons to categorize our existing datasets of SOC texts into pre-defined sentiment value. Systematic and data-driven, these measurement methods ensure that all lexicons faithfully reflects each word's emotionality.

However, this paper does recognize how reductive this quantification of language can be. When it comes to understand such social and human artifacts as language or literary texts, much is dependent on their context. As such, this paper will further discuss these weaknesses of the datasets under Discussion.

2.2 Source Data: Project Gutenberg

Founded in December 1, 1971 by Michael S. Hart, Project Gutenberg exclusively publishes literature in the public domain within the United States. Typically, submissions to Project Gutenberg are digitized editions of printed books, the majority of which were published over 95 years ago. To confirm public domain status, authors can use the copy.pglaf.org website ("Project Gutenberg," n.d.).

The archive relies heavily on volunteers for support and content selection, but it does not accept copyrighted or modern works, even with permission. To submit work, individuals must obtain copyright clearance, scan or capture book images, dedicate hours to proofreading and formatting, and ensure compliance with Project Gutenberg's guidelines, including valid HTML and correct spelling. Common points of failures for submitted works include:

- Crop marks or other printer's marks should not be used on any files.
- The book description should be in compliance with the rules listed here.
- Missing Pages
- Title missing on the front cover
- Incorrect pagination
- Books with typographical errors, such as misspellings or poor grammar.

Most new Project Gutenberg eBooks are in HTML format, which is validated using the W3C's online validator. Additionally, the archive prefers plain text versions whenever possible due to their universal accessibility, and long-term usability. PDF, Word, and other word processor formats are avoided as master formats due to the challenges involved in converting them to HTML and updating them ("Project Gutenberg," n.d.).

With these points being addressed and the impartial nature of the archive being publicly disclosed on its website, the source data seems to archive SOC literary texts with strict quality control and accuracy. However, there are potential spaces for biases such as selection bias where Project Gutenberg primarily hosts older works that have entered the public domain. Thus, archived works might gear towards certain genres, authors, or time periods, potentially limiting the diversity of the datasets for analysis, which will be further discussed below ("Project Gutenberg," n.d.).

2.2.1 Data Cleaning and Word Tokenization

Table 1: An Examplary Table Containing Unprocessed Novel Text (James Joyce)

Book ID	Text	Book	Author
2817	To deep and deeper blue,	Chamber Music	James Joyce
2817		Chamber Music	James Joyce
2817	III At that hour when all things have repose,	|Chamber Music	|James Joyce

To collect text data from five SOC authors using the `gutenbergr` package, we first downloaded the index of each author's works from the Project Gutenberg database. These authors include Joyce, Mansfield, Eliot, Woolf, and Proust (Johnston and Robinson 2023) (Appendix's Table 7 and Table 6).

Next, custom datasets containing each author's raw texts were created. This involved using the identification number of the nine desired literary works and the `gutenberg_download` function to retrieve the HTML texts. Subsequently, the data was saved into individual Parquet datasets, as demonstrated in Table 1.

Table 2: Tokenized Stream of Consciousness Text (James Joyce)

Book ID	Book	Author	Word
4217	A Portrait of the Artist as a Young Man	James Joyce	stead
4217	A Portrait of the Artist as a Young Man	James Joyce	dublin
4217	A Portrait of the Artist as a Young Man	James Joyce	1904
4217	A Portrait of the Artist as a Young Man	James Joyce	trieste
4217	A Portrait of the Artist as a Young Man	James Joyce	1914

Next, in Table 2, to measure the word count of each literary text, we first tokenized or separated each text into individual chunks of words using `unnest_tokens()` (Julia Silge and Robinson 2016). In addition, to account for stop words like, a, of, and, etc., the dataset `stop_words`and the function `anti_join` from Julia Silge and Robinson (2016) were used to exclude all stop words from the self-generated datasets of SOC texts, ensuring data validity and meaningful analysis.

Table 3: Word Count after Tokenization (James Joyce)

Word	Count
stephen	373
god	194
eyes	180
soul	178
father	151

Then, to quantify the number of times each word appears in each text, we used `count()` (Table 3). The result manifests into Figure 1, which visualizes the top 20 words with the highest word count in each SOC author's texts.

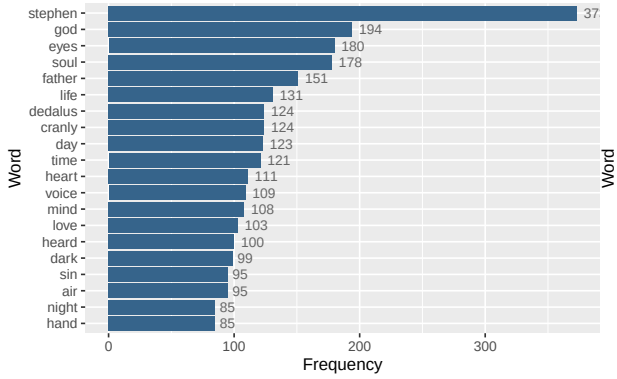
2.2.2 Comparative Word Frequency

Table 4: Word Frequency of Stream of Consciousness Novels, A Comparison Between Five Authors

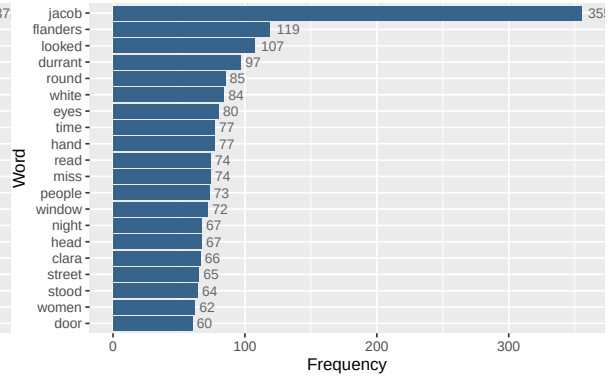
Word	Author	Frequency Percent	Binary Sentiment	Value	Sentiment
accomplished	James Joyce	0.0001120	positive	2	joy
accomplished	James Joyce	0.0001120	positive	2	positive
accomplished	Marcel Proust	0.0000819	positive	2	joy
accomplished	Marcel Proust	0.0000819	positive	2	positive
accomplished	Virginia Woolf	0.0001421	positive	2	joy

To accurately compare word count among SOC authors, from Table 4, we converted each word count from each author into a percentage. This is done by using `mutate` to calculate the word frequency of each word relative to the total word count of each author (`n/sum(n)`). Due to the different number of words present in each author's literary works, this method accounts for the inconsistencies of between each author's word count. The variables within Table 4 are:

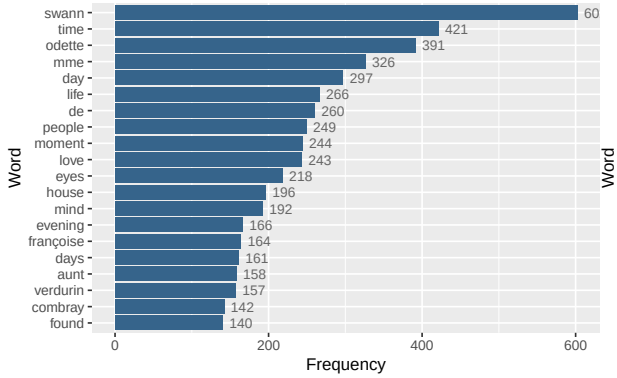
- **Word:** This variable is the word sourced from SOC literary texts.
- **Author:** This variable accounts for the author whose books contain the word in question.
- **Frequency:** This variable is the word count percentage of the word in question against each author's total word count.
- **Binary Sentiment (Bing):** This variable, as explained, comes from the Bing sentiment index, enabling us to analyze how every word is usually perceived emotionally, whether negatively or positively.
- **Value of Sentiment (AFINN):** This variable, as explained, comes from the AFINN sentiment index, enabling us to analyze how every word is usually perceived emotionally, on a scale between -5 and 5.
- **Sentiment (NRC):** This variable, as explained, comes from the NRC sentiment index, enabling us to analyze how every word is usually perceived emotionally, whether it's associated with anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise or trust.



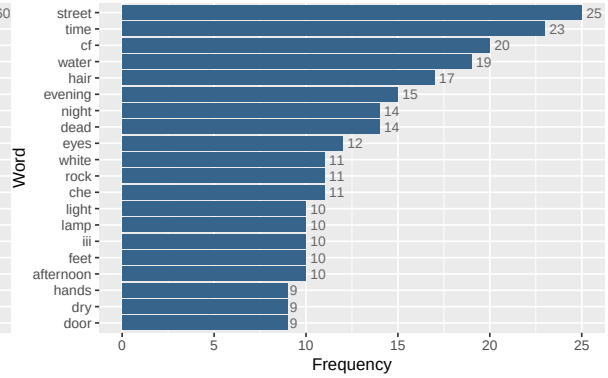
(a) James Joyce



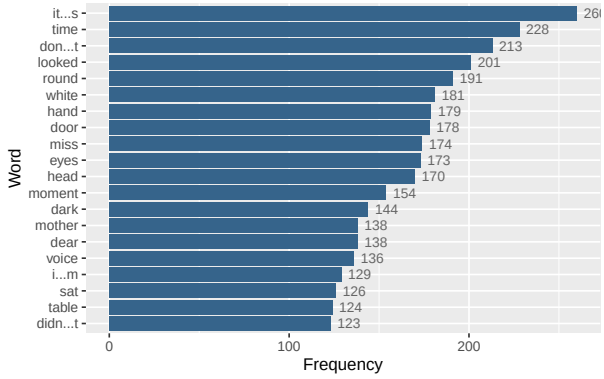
(b) Virginia Woolf



(c) Marcel Proust



(d) T.S. Eliot



(e) Katherine Mansfield

Figure 1: Comparative Analysis of Top 20 Word Frequencies from Famous SOC Authors

Table 5: Overall Boxplot Summary Statistics of all Word Frequency Percentages (AFINN Index)

Frequency	Min	Q1	Median	Mean	Q3	Max	SD
0.0000819	-4	-2.00	-2	-0.5550756	2	5	2.2802619
0.0001120	-5	-2.00	-2	-0.6313181	2	4	2.3035623
0.0001277	-5	-2.00	-2	-0.5358696	2	5	2.4156761
0.0001421	-4	-2.00	-2	-0.4638243	2	4	2.2845697
0.0002240	-4	-4.00	-2	-0.7692308	2	3	3.0863637
0.0002555	-3	-3.00	-2	-0.5769231	2	3	2.4523144
0.0004255	-3	-2.00	-2	-0.5733333	2	4	2.2331112
0.0008511	-3	-2.75	-2	-2.3333333	-2	-2	0.5163978

Table 5 captures the text data garnered and their emotional connotations using the AFINN lexicon. Overseeing the entire dataset, Table 5 demonstrates that, for all distinct word frequency, the median sentiment value is -2 on the scale between -5 and 5. More specifically, words that appear most frequently, or words with the frequency of 0.0008511%, have the most negative connotation of -2.33 on the scale between -5 and 5 in the AFINN sentiment index. This summary statistic further shows how SOC literature mostly associates with words that carry more negative or even extreme connotations often associated with mental health themes, and topics such as depression, disassociation or schizophrenia (Nyongesa 2023), a central argument that we will further discuss in Results.

2.3 Data Limitations

2.3.1 Potential Measurement Errors

The actual raw text data sourced from Project Gutenberg may be susceptible to measurement errors due to various factors, such as selection bias towards older, Western and canonical texts, social desirability bias (which edits out certain political elements), memory lapses, misinterpretations of literary texts by voluntary proofreaders. While the archive has strict guidelines to minimize these errors, they cannot be entirely eliminated.

2.3.2 Uneven Text Length

We chose literary texts based on their global reception and status within the stream-of-consciousness genre. This decision didn't consider certain practical constraints, such as ensuring similar text lengths across all works by each author. While this approach enables us to thoroughly explore word frequency analysis, sentiment value, and word networks within the SOC genre, it also has a downside: privileging the frequencies of certain words from specific

authors more than others due to the longer text length of those authors. This could critically affect the integrity of the data analysis.

2.3.3 Biased Author Selection

Furthermore, SOC literature is expansive, spanning a wide range of themes, styles, and narrative techniques. By exclusively representing the genre through a select group of canonical Western authors, we risk constraining the genre's diversity. By oversimplifying the complexity of the literary genre, we also affect the integrity of the datasets and its correlation to mental health themes.

3 Results

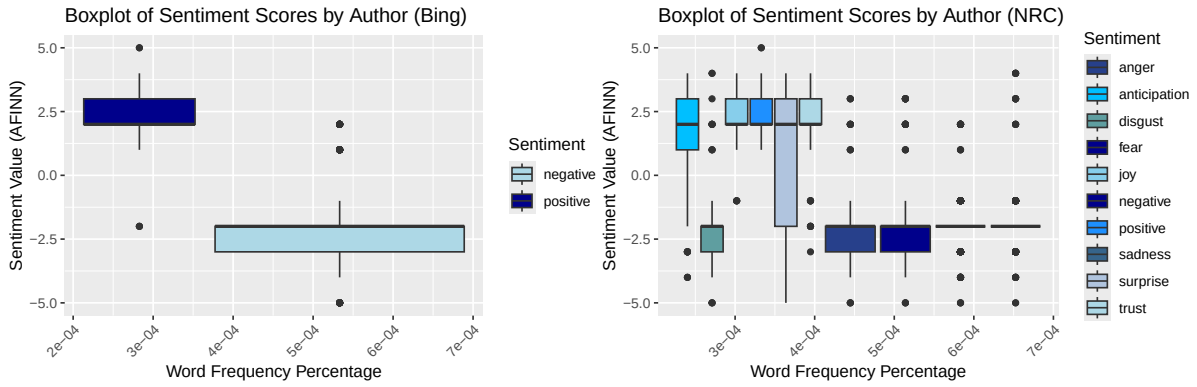
Upon assessing our generated datasets from Project Gutenberg, we further extrapolate on them here by combining word count with the AFINN, Bing and NRC index to conduct relevant word frequency and sentiment analysis. The later sections of Results also compound on these findings to conduct a comparative analysis between gendered SOC authors, attempting to garner relevant insights in mental health themes in different demographics of the SOC genre.

3.1 The Dominant Vocabulary of SOC Literature, A Word Frequency Analysis

In terms of word frequency, most words depicted in the nine selected SOC texts are abstract nouns with temporal or philosophical tendencies and negative connotations.

Figure 1 demonstrates that beside personal names, words with the highest counts are mostly abstract nouns that revolve around the nature of time, and life. For instance, these words are, *god* (appearing 194 times in Joyce's), *time* (appearing in 4 authors' texts, 121 times for Joyce's, 77 times for Woolf's, 421 times for Proust's, 23 times for Eliot), *mind* (108 times in Joyce's and 196 times in Proust's), *life* (appearing 266 times in Proust's and 131 times in Joyce's), etc. This further exemplifies the sense of disassociation, temporality and non-linearity commonly associated to stream-of-consciousness literature and more broadly with mental health themes (Long and So 2016) (Nyongesa 2023) (Bernini and Fernyhough 2022).

Additionally, within Figure 2, for the NRC index, between all authors, the words with lower frequency-those ranging from 0.0000819% to 0004255%-are largely words that carry a mixture of emotional connotations such as, joy, surprise, positive, disgust and anticipation. However, words with high frequency-ranging from 0,0004255% to 0.0008511%-mostly contain negative connotations like fear and negative. They also have a negative -1.875 median on a scale between -5 to 5 in the AFINN index. Similarly, for the Bing index, words with higher frequency-those ranging from 0,0004255% to 0.0008511%-largely carry negative connotations. They also have, approximately, a -1.875 median on a scale between -5 to 5 in the AFINN index. This further



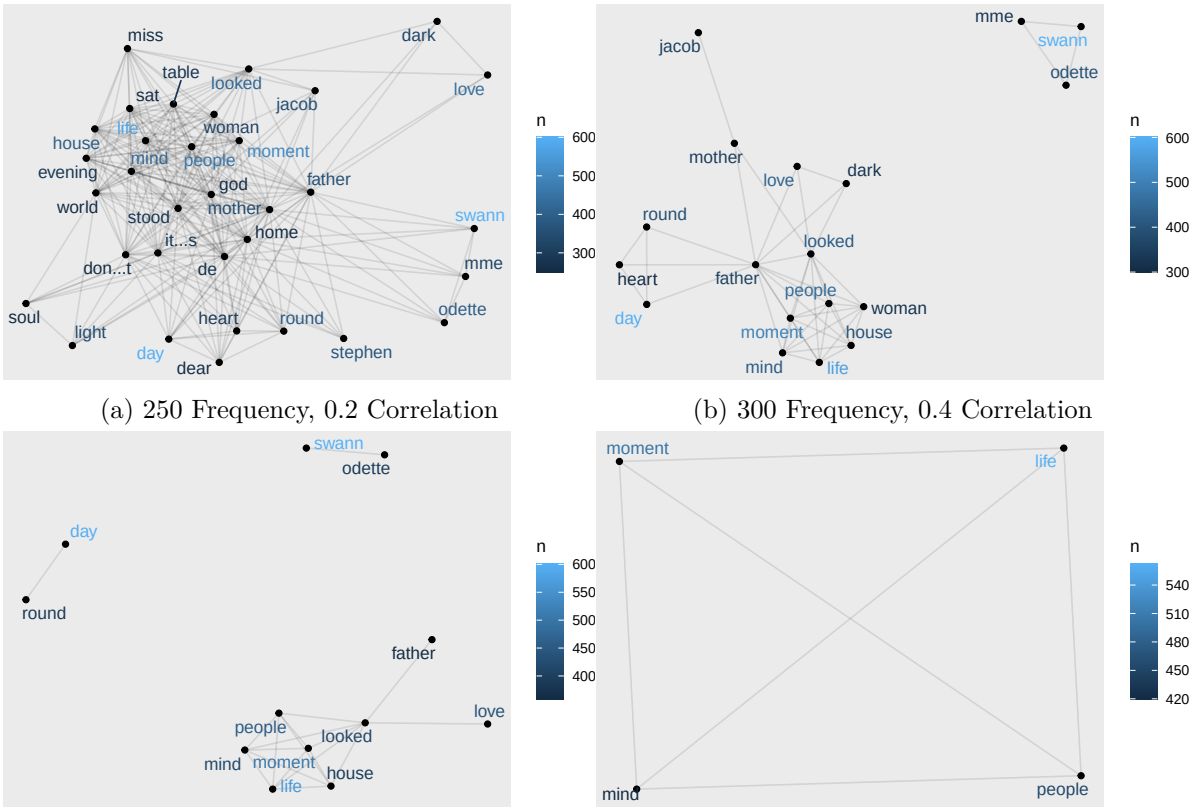
(a) Bing lexicon (b) NRC lexicon

Figure 2: Overall Word Frequency and Sentiment Analysis

emphasizes how SOC literature mostly contains negativity-inclined vocabulary and linguistic patterns.

To further explore the contextual relationship between each word, Figure 3 creates word networks of all nine SOC texts. These networks capture how often each word frequents in these SOC texts and how closely each correlates with one another, whether as single words (unigrams) or pairs of words (bigrams). When combined all SOC texts, temporal nouns like *time, day, life, moment, night* appear the most, at, respectively, 870, 597, 567, 502 and 376 times (Appendix's Table 8). This further correlates to the non-linear nature of SOC literature and their exploration of cognitive incoherence. In terms of correlation, ignoring personal names like *swann, odette or mme*, bigrams with a correlation of 1 include word pairs like *dark, love*; *day, heart*; *moment, life*; *mind, life*; *moment, mind*, among others (Appendix's Table 9). Observable from Figure 3, with a correlation and a word count of at least 0.8 and 400, *people, moment, life* and *mind* are the most predominant word combinations throughout all 9 SOC texts.

This word frequency and correlation analysis then has enabled us to grasp the thematic framework of SOC literature, where most correlated and frequently present words are those with philosophical undertones, a sense of temporality and tinge of mental incoherence like *dark, love*. While not entirely direct, this has somewhat portrayed SOC literature as being intricately linked to cognitive dissonance or the mind, and especially mental health issues in which SOC authors have portrayed through their characters (Long and So 2016) (Nyongesa 2023) (Bernini and Fernyhough 2022).



(a) 250 Frequency, 0.2 Correlation

(b) 300 Frequency, 0.4 Correlation

(c) 350 Frequency, 0.6 Correlation

(d) 400 Frequency, 0.8 Correlation

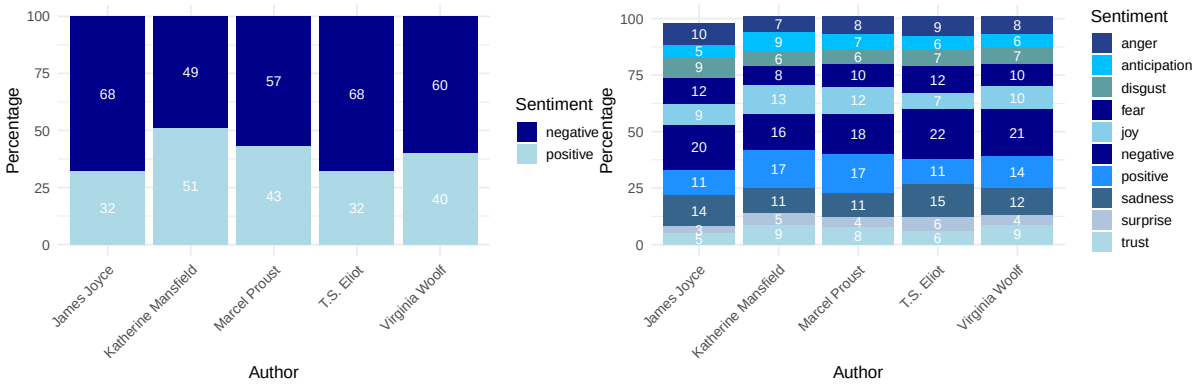
Figure 3: Word Networks Measured by Frequency and Correlation when Combining All Stream of Consciousness Novels

3.2 Sentiment Analysis

To further understand the emotional landscape and mental health themes related to SOC literature, all nine SOC texts are combined to conduct an overall sentiment analysis (Mohammad and Turney 2013).

For the Bing index, Figure 4 shows that, SOC authors' literary works are predominantly negative. All authors, except for Mansfield, demonstrates this. 68% of Joyce's works are negative compared to his 32% positive, similar to Eliot's, which is also the highest statistic out of all five authors. Woolf comes second with 60% words carrying negative meanings, while 40% are negative. Proust comes fourth with 57% negative and 43% positive. While Mansfield has words with positive connotations, at 51%, this isn't much different from her 49% negative words.

For the NRC index, Figure 4 further details the emotional nuances of SOC literature. The dominant emotional connotation in SOC literature's vocabulary is negative, standing at 16% to 22%. While the second highest percentage range belongs to words with positive connotation, ranging from 11% to 17%. However, most of the other bars in the graph are occupied by words with negative-leaning emotions, such as sadness, fear, and anger. Sadness-related words, in particular, range the third highest, from 11% to 15%.



(a) Bing lexicon

(b) NRC lexicon

Figure 4: Categorical Sentiment Analysis of All SOC Novel Texts by Authors

The AFINN index in Figure 5 portrays a similar trend. On a scale of -5 to 5, SOC literature from these five authors mostly carry a negative sentiment, observable in the significant difference in height between the negative trendlines and the positive trendlines. Eliot's works are the most negative with 37% of words being -2 on the emotional scale between -5 and 5. Respectively, Woolf's, Proust's and Joyce's tail beind at 33%, 32% and 31%, averaging out to be -2.2 on the emotional scale between -5 and 5. While a great portion of Mansfield's works are positive, 24% of her words-her largest percentage-are negative, earning a -2.2 negative score between -5 and 5 in the AFINN index.

This further suggests that the SOC literary landscape mainly has negative, disgust, fear and sadness-leaning vocabulary. These trends reflect the incoherent mental spaces, depressive perceptions, and existential angst inherent to SOC literature. These are the central themes often conveyed by SOC authors through their exploration of the unconscious mind and repressed emotions, reflecting the societal concerns of the late 19th to mid-20th century West (Nyongesa 2023) (Bernini and Fernyhough 2022).

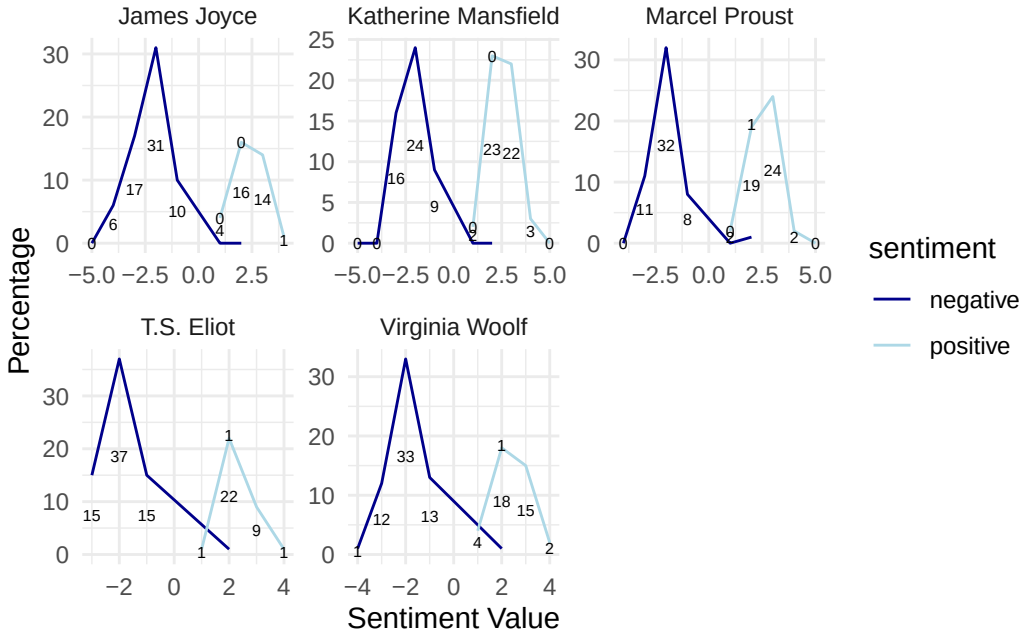


Figure 5: Numerical Sentiment Analysis of All SOC Novel Texts by Authors (AFINN lexicon)

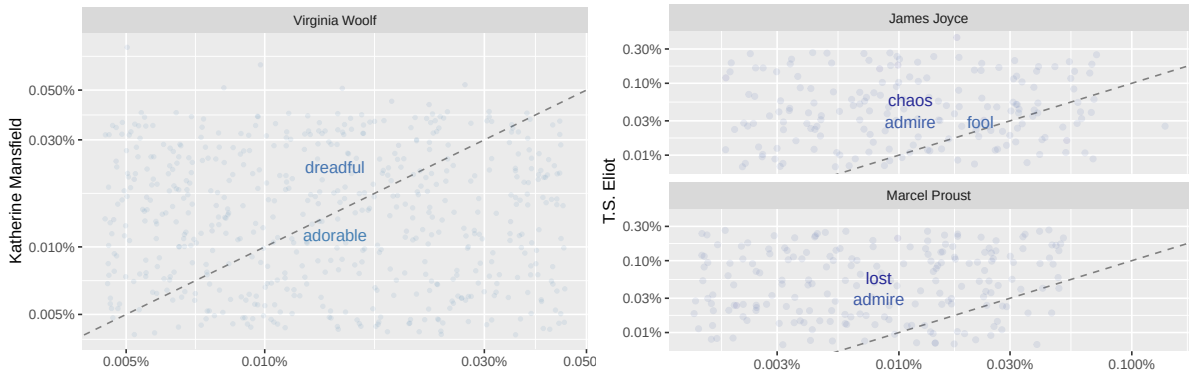
3.3 Gendered Mental Landscape of Stream of Consciousness Authorw

To further expand on the analysis, we have also used our word frequency data from Table 4 to conduct a gendered analysis between different SOC authors. Figure 6 attempts to assess the word frequency between female and male SOC authors. Dots closer together around the diagonal line tend to appear with similar frequency for all authors involved. Words further up and to the left are more typical for those on the y-axis, and closer to the lower right corner: those on the x-axis.

Figure 6 shows that for female SOC authors, while having smaller word frequency due to the lighter blue, still have their words perfectly revolved the regression line, showing that both Woolf and Mansfield are comparatively close in terms of linguistic patterns. Woolf and Mansfield mostly have a word frequency of 0.0001421262% and 0.0001277302% for each word present in their texts, with some of their shared vocabulary being *anxious*, *angry*, *anguish*, *dreadful*, among others.

As for male authors, Figure 6 shows that, while their words have a higher appearance frequency, they have less correlation as the regression line isn't centered, showing the dissimilar vocabulary used between male authors. Mostly, Eliot, Proust and Joyce have a word frequency of 0.0004255319%, 0.0000819% and 0.000112% for each word present in their texts, with some of their shared vocabulary being *agony*, *cry*, *die*, *chaos*, among others. Proust and Joyce, two novelists, seem to have more similarities compared to Eliot, who is a poet.

Through this gendered analysis and Figure 4, we can see discrepancies between male and female SOC authors, especially in terms of shared linguistic patterns and mental health themes. Lettieri and Cecchetti (2023) poses that in terms of genders, writers-or here SOC writers-tend to write about affects or mental health themes differently. While females tend to have reservations over their emotions or are similarly positive, males tend to be more open and individualistic due to less social constraints at the time, which will be further discussed in Discussion below.



(a) Female Authors

(b) Male Authors

Figure 6: Comparative Analysis of Word Frequency in Female and Male Stream of Consciousness Authors

4 Discussion

4.1 Mental Health Vocabulary: Patterns and Trends

The findings in this paper reveals a correlative relationship between SOC literature and mental health-inclined vocabulary and linguistic patterns. In terms of frequency, a dominant portion-or over 60% of words in the NRC index-of select SOC texts in this analysis have negative and sadness-leaning words (Figure 2) (Figure 4) (Figure 5). Most repeated words are abstract, temporal nouns like *time*, *mind* or *life*, which are aligned with SOC literature's exploration of mind wandering, non-linearity and unconscious cognition (Figure 3).

As scholars Bernini and Fernyhough mentioned, the "mysterious nature of the conscious" is "a simultaneously fragmented and unified mental realm" (Bernini and Fernyhough 2022), this analysis has also reflected that sentiment in SOC literature. Most of the texts express the desire to textually replicate the inner workings of the human mind, attempting to capture the complexity of thought processes and emotions. With these inner-workings revealed to be mostly negative and existential, SOC authors have explored their identity and worldview in an authentic manner, deeply portraying then-taboo topics such as existential angst, anxiety and cognitive dissonance.

4.2 Insights into Socio-Political Landscape of the West's Modernist Era

Through the analysis, we can see the SOC texts' linguistic patterns reflect the socio-political landscape of Western society. The SOC genre originated from modernist literature, whose writers often rebelled against linear and traditional narrative techniques of the 19th century (Bernini and Fernyhough 2022). According to scholar Nyongesa, "marginalization occasioned by the dominant group result in neurotic conditions in members of the minority group" (Nyongesa 2023), which rings true to the scenario at hand about SOC literature and its relationship to its birthplace: late 19th to mid-20th century Western society.

There are several reasons rooted in the West's modernist era that resulted in the growth and popularity of the SOC genre. This period's rapid industrialization and urbanization reshaped the traditional institutions of Western society, leading to fragmentation and dislocation, which SOC authors had attempted to capture in their own works (Long and So 2016). Advancing transportation, photographic media and mass communication affected the West's sense of time and space, which in turn affected how people understood reality. This is a central pillar of SOC literature: distorted reality, atemporality and an unconscious perception of the world.

This time period also saw the West's crisis of faith in traditional religious beliefs. With the emergence of Darwin's theory of evolution, Freudian psychology, and existential philosophy came different notions of human existence, morality, and meaning (Bernini and Fernyhough 2022). The heightened search for individuality had SOC authors embraced themes such as alienation, identity and mental health in an increasing fragmented world. The devastation from World War I and the interwar period also further intensified this crisis. Post-war collapse of nations created new ideologies like communism and fascism. This in turn created an atmosphere of disillusionment and uncertainty that SOC literature attempted to answer with its anxious and disassociative writing style.

4.3 Gender Differences within Stream of Consciousness Genre and Mental Health

Further, our gendered analysis of SOC literature in Figure 6 reveals differences in linguistic patterns between male and female authors during this time period, particularly in their portrayal of affects and mental health themes. Female writers tend to exhibit more reserved

emotions or positivity, while male writers display greater emotional range and individualism, influenced by restrictions of female expressions at their time.

Research findings indicate that, on average, female writers use positive words more frequently than male writers, reflecting societal norms attributing positive emotions to females (Lettieri and Cecchetti 2023). The similarity in vocabulary among SOC female writers like Woolf and Mansfield may stem from shared-or restricted-literary influences, higher cultural and social expectations, common personal experiences, and considerations of reader norms during the West's late 19th to mid-20th century. These factors contribute to a nuanced understanding of gendered linguistic patterns, which has been shown to be more forgiving towards male authors than female authors in the SOC genre.

4.4 Weaknesses

4.4.1 Lack of Thorough Word Cleaning

With the large amount of different literary texts need processing, it's hard to truly conduct thorough word processing. Certain texts, while can have their frequencies accounted for, can't have sentiment analysis due to their factual nature. These textual elements might include things like chapter titles, texts from a language other than English, Roman numbers and personal names, which can strongly skew and affect the data integrity. As observed in Figure 1, *Swann*, *Stephen* or *Jacob* are the most repeated words in Proust's, Joyce's and Woolf's, proving this data limitation.

4.4.2 Decontextualized Words and Limiting Literature Selection

Throughout our analysis, words are singled out and analyzed without contexts which could have affected their intended meanings, especially in such a complex genre such as SOC. For instance, the commonly repeated words *time*, *mind* or *life* in our selected SOC texts might mean something entirely different when combined with other words. Paradoxically, these words may be insignificant to the SOC texts in question themselves as they're only small units of already expansive literary works that often have complex themes. Therefore, word frequency analysis might have not been nuanced when analyzing the contexts behind mental health themes and their relationship to SOC literature. In addition, not all SOC texts in the archive are available to be transcribed into datasets due to licensing issues, further limiting our analysis.

4.4.3 Reliance on A Small Number of Sentiment Indices

To ensure a thorough sentiment analysis of SOC texts, this paper has leveraged three widely used sentiment indices, namely Bing, NRC and AFINN. However, a sole dependence on the emotional metrics of these three indices might not be nuanced enough. This is because to

categorize words into a specific emotion is an arduous and complex tasks, as emotions aren't easily quantifiable and words are mainly context-based. In addition, some of the indices used might not ensure valid data analysis. For instance, Table 4 shows that, using the NRC index, the word *accomplished* can mean two different emotions: joy and positive. This flexible categorization can very much skew the results of the sentiment analysis, bringing one emotion more to the forefront than others.

4.5 Moving Forward and Next Steps

Moving forward, to gain a deeper understanding of the relationship between SOC literature and mental health themes, more diverse and better natural language processing (NLP) indices, methods and processes must be used to conduct a more accurate analysis.

As mentioned, relying on three sentiment indices are not enough to accurately capture the emotional sentiments of the words used in SOC literature due to how context-based words are. More relevant indices then should be incorporated into the analysis.

In addition, due to the abstract nature of SOC texts, using advanced NLP models is more effective for analysis rather than focusing solely on small linguistic units such as words or word frequency. Topic modeling, for example, is a technique in NLP used to uncover abstract topics within texts like SOC literature. Leveraging algorithms like Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), topic modeling helps identify patterns of co-occurring words within SOC literature, grouping them into clusters that represent coherent themes or topics that can be relevant to our mental health analysis of the SOC genre.

Lastly, due to the qualitative nature of SOC literature, using a more mixed methods approach-combining both quantitative and qualitative data analysis-will offer a more comprehensive understanding of the broader mental health trends and the literary genre itself.

5 Conclusion

This study analyzes stream of consciousness (SOC) literature's linguistic patterns, revealing its portrayal of mental health themes. Through our analysis of works by Joyce, Woolf, Proust, Mansfield, and Eliot, we uncover that SOC literature is largely negative or sadness-inclined with a philosophical and temporal undertones due to its exploration of unconscious cognition. In addition, expression and sentiment-wise, female authors' writing is more restricted compared to male authors, revealing gendered differences in SOC literature. While our quantitative analysis provides insights into the genre's relationship with mental health themes, it's important to recognize its limitations in capturing the richness of SOC language. Therefore, moving forward, using more advanced NLP methods will better benefit our exploration of the relationship between SOC literature and mental health themes. We hypothesize that this can

provide deeper insights into the genre, and the reasons for its growth in Western society during late 19th to mid-20th century.

6 Appendix

6.1 All Books from Project Gutenberg Archive

Table 6: A Sample of Project Gutenberg Archive's Book Database

Book ID	Title	Author	Text Included
6	Give Me Liberty or Give Me Death	Henry, Patrick	TRUE
7	The Mayflower Compact	NA	TRUE
8	Abraham Lincoln's Second Inaugural Address	Lincoln, Abraham	TRUE
9	Abraham Lincoln's First Inaugural Address	Lincoln, Abraham	TRUE
10	The King James Version of the Bible	NA	TRUE

6.2 Combined Bibliographies of Nine Selected Stream of Consciousness Texts

Table 7: Sample Combined Bibliographies of 9 Selected Stream of Consciousness Texts

Book ID	Title	Author	Text Included
2814	Dubliners	Joyce, James	FALSE
2817	Chamber Music	Joyce, James	FALSE
4217	A Portrait of the Artist as a Young Man	Joyce, James	FALSE
4300	Ulysses	Joyce, James	FALSE
7872	Dubliners	Joyce, James	FALSE

6.3 Combined Statistcs

6.3.1 Combined Count

Table 8: Sample Combined Count of Words from All Nine Stream of Consciousness Texts

Word	Count
time	870
eyes	663
swann	602
day	597
life	563

6.3.2 Combined Correlation

Table 9: Sample Combined Bibliographies of 9 Selected Stream of Consciousness Texts

Word 1	Word 2	Correlation
moment	people	1
mind	people	1
life	people	1
people	moment	1
mind	moment	1

References

Allaire, J., Y. Xie, C. Dervieux, J. McPherson, J. Luraschi, K. Ushey, A. Atkins, et al. 2024. *Rmarkdown: Dynamic Documents for r*. R package version 2.26. https://github.com/rstudio/rmarkdown.

Bernini, M., and C. Fernyhough. 2022. "Resampling (Narrative) Stream of Consciousness: Mind Wandering, Inner Speech, and Reading as Reversed Introspection." *Modern Fiction Studies* 68 (4): 639–67. https://doi.org/https://doi.org/10.1353/mfs.2022.0045.

Feinerer, I., K. Hornik, and D. Meyer. 2008. "Text Mining Infrastructure in r." *Journal of Statistical Software* 25 (5): 1–54. https://doi.org/10.18637/jss.v025.i05.

Hvitfeldt, Emil. 2022. *Textdata: Download and Load Various Text Datasets*. https://github.com/EmilHvitfeldt/textdata.

Johnston, Myfanwy, and David Robinson. 2023. *Gutenbergr: Download and Process Public Domain Works from Project Gutenberg*. https://docs.ropensci.org/gutenbergr/.

Lettieri, Handjaras, G., and L. Cecchetti. 2023. "How Male and Female Literary Authors Write about Affect Across Cultures and over Historical Periods." *Affective Science* 4 (4): 770–80. https://doi.org/10.1007/s42761-023-00219-9.

Long, H., and J. So R. 2016. "Turbulent Flow: A Computational Model of World Literature." *Modern Language Quarterly* 77 (3): 345–67. https://doi.org/https://doi.org/10.1215/00267929-3570656.

Mohammad, Saif M., and Peter D. Turney. 2013. "Crowdsourcing a Word-Emotion Association Lexicon." *Computational Intelligence* 29 (3): 436–65. https://doi.org/10.1111/j.1467-8640.2012.00460.x.

Müller, Kirill. 2020b. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

———. 2020a. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Nyongesa, A. 2023. "The Centre and Pathology: Postmodernist Reading of Madness in the Oppressor in Contemporary Fiction." *Cogent Arts & Humanities* 10 (1): 1–12. https://doi.org/https://doi.org/10.1080/23311983.2023.2249280.

Pedersen, L., T. 2024. *Ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. https://ggraph.data-imaginist.com.

"Project Gutenberg." n.d. https://www.gutenberg.org.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Silge, J., and D. Robinson. 2006. "The Igraph Software Package for Complex Network Research." *InterJournal,*Complex Systems**, 1695. https://igraph.org.

———. 2022. *Widyr: Widen, Process, Then Re-Tidy Data*. https://github.com/juliasilge/wi

dyr.

Silge, Julia, and David Robinson. 2016. "Tidytext: Text Mining and Analysis Using Tidy Data Principles in r." *JOSS* 1 (3). https://doi.org/10.21105/joss.00037.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. *Scales: Scale Functions for Visualization*. https://scales.r-lib.org.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.