# 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 22, 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, this paper aims to uncover shared linguistic patterns and gain insights into the authors' mental states. This paper attempts 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. (add one sentence on main results)

## 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 inner workings of the character's mind in an intimate and unfiltered manner (Long and So 2016). In literature, most scholars agree that stream of consciousness reveals the complexities of mind-scapes, shedding light on the nuances of characters' emotional well-being and psychological struggles (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, I seek 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 geographical locations and 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 disabilities.

Thus, the estimand is the correlation between mental health-related themes and words in SOC literature, their frequency and sentiment value as provided by Julia Silge and Robinson (2016). This is considered in terms of nine selected SOC novels 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 (percentage, number and data here, main results)...

To further understand the correlation between stream of consciousness novels and mental health themes, in Introduction, the paper briefly discusses the nature of stream of consciousness literature, relevant authors and the works that I've chosen to analyze. Subsequently, in Data and Results, I 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. To complete the paper, Appendix clarifies how each variable within each dataset is generated with relevant tables to accordingly demonstrate this.

The novel 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).

# 2 Data

## 2.1 Measurement

Two central variables in this paper are:

- **Word Frequency:** This variable captures the repetition of a single word (unigram), two-words combination (bigram) or three-words combination (trigram) 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` 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, I first downloaded all nine novel texts from "Project Gutenberg" (n.d.), and leveraged functions such as `unnest_tokens()` from Julia Silge and Robinson (2016) to mine the texts, or separating it into individual words. Finally, I used `count()` to quantify the word frequency.

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 my sentiment analysis are (Julia Silge and Robinson 2016):

- 'AFINN' from Finn Årup Nielsen, which assigns a numerical value from '-5 to 5'
- 'bing' from Bing Liu and collaborators, which assigns if a word is '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")` in which I can use `inner_join` to categorize my existing datasets of novel texts into their sentiment value. Systematic and data-driven, these measurement methods ensure that all lexicons faithfully reflects each word's emotionality.

However, I do 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, I 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 in which I garner my literary texts are selected with strict quality 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 | NA | Chamber Music | James Joyce |
| 2817 | III At that hour when all things have repose, | \|Chamber Music | \|James Joyce |

To garner the word count of SOC literature using the library `gutenbergr`, I first downloaded the index of each respective SOC author's body of works using each author's name as outlined in Project Gutenberg database, namely Joyce, Mansfield, Eliot, Woolf, and Proust (Johnston and Robinson 2023) (Appendix). Then, to create my own datasets containing the word frequency within each SOC author's texts, I used the identification number of the desired literary work and the function `gutenberg_download` to upload the HTML texts into individual csv-and later on, parquet-datasets, as observed in Table 1 for James Joyce.

Table 2: Tokenized Novel Text & Word Count after Tokenization (James Joyce)

| Book ID | Author | Word |
|---|---|---|
| 4217 | James Joyce | 1904 |
| 4217 | James Joyce | trieste |
| 4217 | James Joyce | 1914 |

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

Next, observable in Table 2, to measure the word frequency of each literary text, I first tokenized or separated the texts 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., I leveraged the dataset `stop_words`and the function `anti_join` from Julia Silge and Robinson (2016) to exclude all stop words from my self-generated datasets of literary texts, ensuring data validity and meaningful analysis.

Then, to quantify the number of times each word appears in each text, I used `count()` (Table 2). The result manifests into Figure 1. The figure demonstrates that beside the top frequencies of personal names, most of the words are abstract nouns that revolve around the nature of time, and life, e.g. 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).

### 2.2.2 Comparative Word Frequency

Table 3: Word Frequency of Stream of Consciousness Novels, A Comparison Between Five Authors

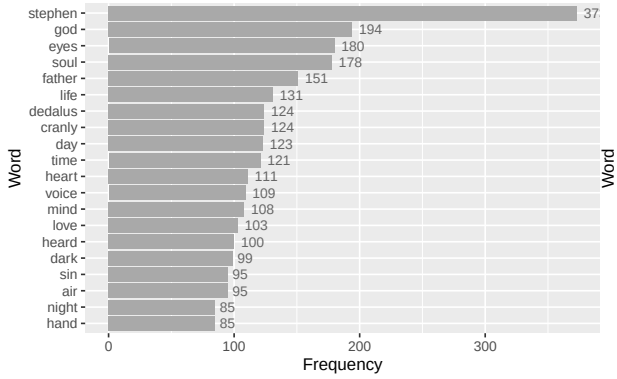
| Word | James Joyce | Katherine Mansfield | Marcel Proust | T.S. Eliot | Virignia Woolf |
|---|---|---|---|---|---|
| abandon | 0.000112 | NA | 8.19e-05 | NA | NA |
| abandoned | 0.000112 | NA | 8.19e-05 | NA | NA |
| abandonment | 0.000112 | NA | 8.19e-05 | NA | 0.0001421 |
| abase | 0.000112 | NA | NA | NA | NA |
| abased | 0.000112 | NA | NA | NA | NA |

To accurately compare word frequency among SOC authors, from Table 3, I convert each word frequency into a percentage. This is done by using `mutate` to calculate the word frequency of each word relative to the total word frequency of each author. 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.

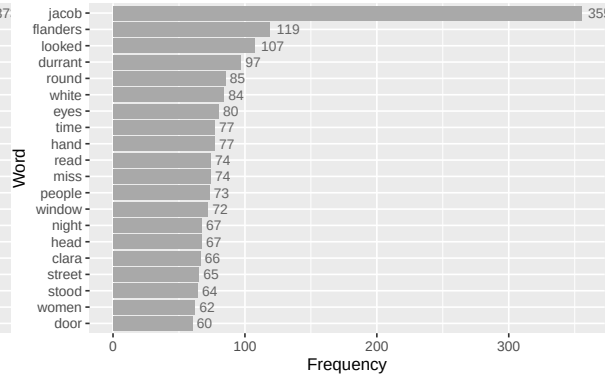
Table 4: Overall Boxplot Summary Statistics of all Word Frequency Percentages (AFINN Sentiment)

| 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 |

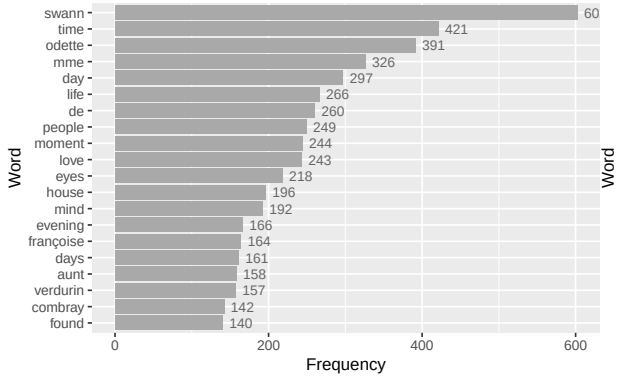
Figure 2 and Table 4 captures the data garnered. For the NRC index, for all authors, the lower word frequency percentage ranges from 0.0000819 to 0004255 and are largely words that carry a mixture of emotional connotations such as, joy, surprise, positive, disgust and anticipation. However, the higher word frequency percentage, ranging from 0,0004255 to 0.0008511, mostly contains words with negative connotations like fear and anger, that weighs a -1.875 median on a scale between -5 to 5 in the AFINN index. Similarly, for the Bing index, the higher word frequency ranging from 0,0004255 to 0.0008511 percentages largely contains words that carry negative connotations, that weighs approximately a -1.875 median on a scale between -5 to 5 in the AFINN index.



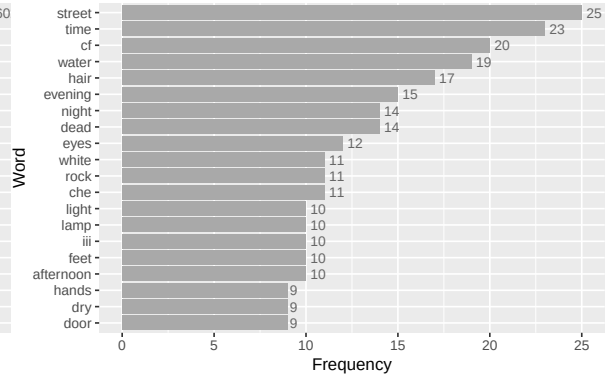

(a) James Joyce



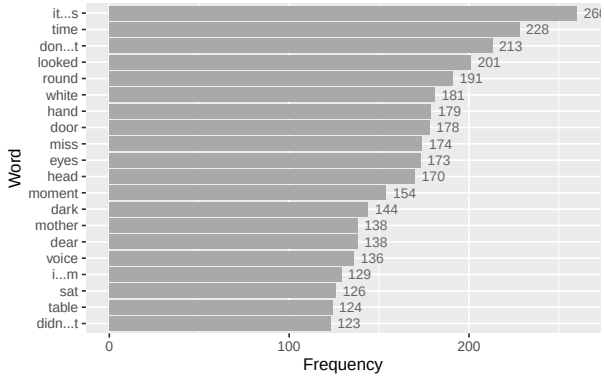

(b) Virginia Woolf




(c) Marcel Proust



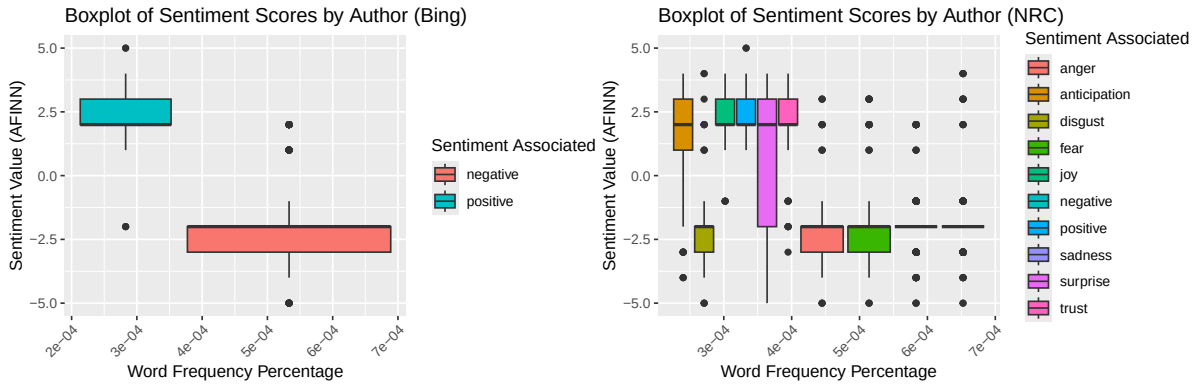

(d) T.S. Eliot




(e) Katherine Mansfield

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

Overseeing the entire dataset, Table 4 demonstrates that, for all distinct word frequency percentages, 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 percentage of 0.0008511, have the lowest mean sentiment value-or the most negative connotation of-2.33 on the scale between -5 and 5 in the AFINN sentiment index. This summary statistics of the gathered data only further shows how the SOC literary landscape 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) Bing lexicon

(b) NRC lexicon

Figure 2: Overall Word Frequency and Sentiment Analysis

## 2.3 Data Limitations

### 2.3.1 Lack of Thorough Word Cleaning

With the large amount of different literary texts need processing by myself, it's hard to truly conduct thorough word processing. Certain textual elements, 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, 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.

### 2.3.2 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 that edits out certain political elements, memory lapses, or misinterpretations of literary texts by voluntary proofreaders. While the archive has strict guidelines to minimize these errors, they cannot be entirely eliminated.

#### 2.3.3 Uneven Text Length

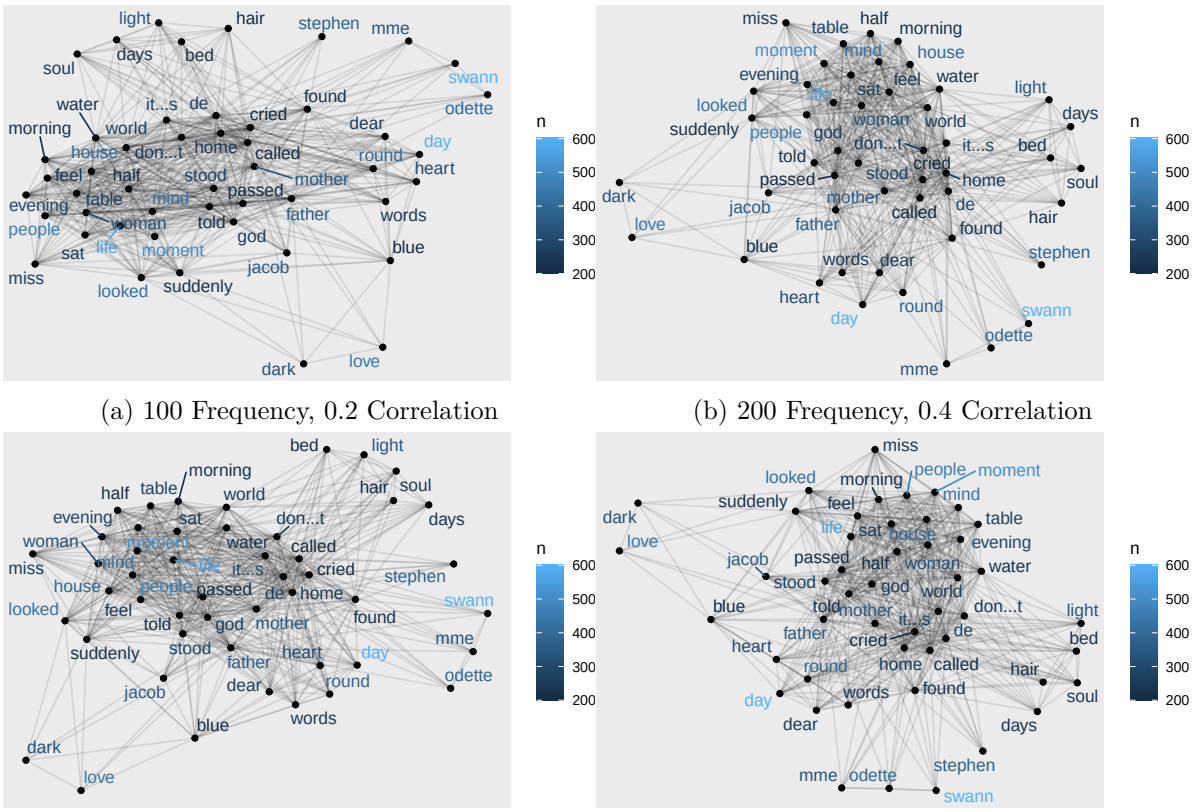
I chose literary texts based on their global reception and canonical status within the stream-of-consciousness genre. This decision involved forgoing certain practical considerations, such as ensuring similar text lengths across all works by each author. While this approach enables a thorough exploration of word frequency analysis, sentiment value, and word networks within the SOC genre, it also has the potential to emphasize 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.

### 2.4 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, I risk constraining the genre's diversity. By oversimplifying the complexity of the the dynamic literary genre, I also affect the integrity of the datasets and its correlation to mental health themes.

## 3 Results

### 3.1 The Dominant Vocabulary of SOC Literature




(a) 100 Frequency, 0.2 Correlation

(b) 200 Frequency, 0.4 Correlation

(c) 200 Frequency, 0.8 Correlation

(d) 400 Frequency, 1 Correlation

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

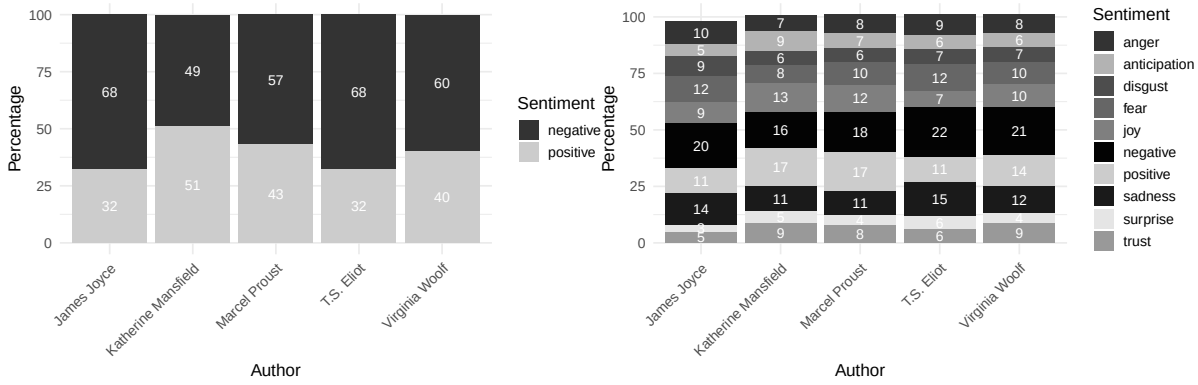
## 3.2 Sentiment Analysis

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

For 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.

The NRC index in 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 belong to words with positive connotation, ranging from 11% to 17%, most of the other bars are occupied by words with negative-leaning emotions, such as sadness, fear, and anger. Sadness-related words, in particular, ranges 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 with its more quantitative sentiment analysis. 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, Proust and Joyce tail beind at 33%, 32% and 31%, averaging out to be ~-2.2 on the emotional scale between -5 and 5. While most a large percentage Mansfield's works are positive, her highest percentage is negative on the AFINN index, at 24% for a scale of about -2.2 between -5 and 5.

This further suggests that the SOC literary landscape is inundated with negative and sadness-leaning vocabulary. These trends portray the the depressive or existential angst themes often appearing in the genre, especially when a lot of these SOC authors dealt with the unconscious minds and repressed emotions within the West's late 19th to mid-20th century society (Nyongesa 2023) (Bernini and Fernyhough 2022).

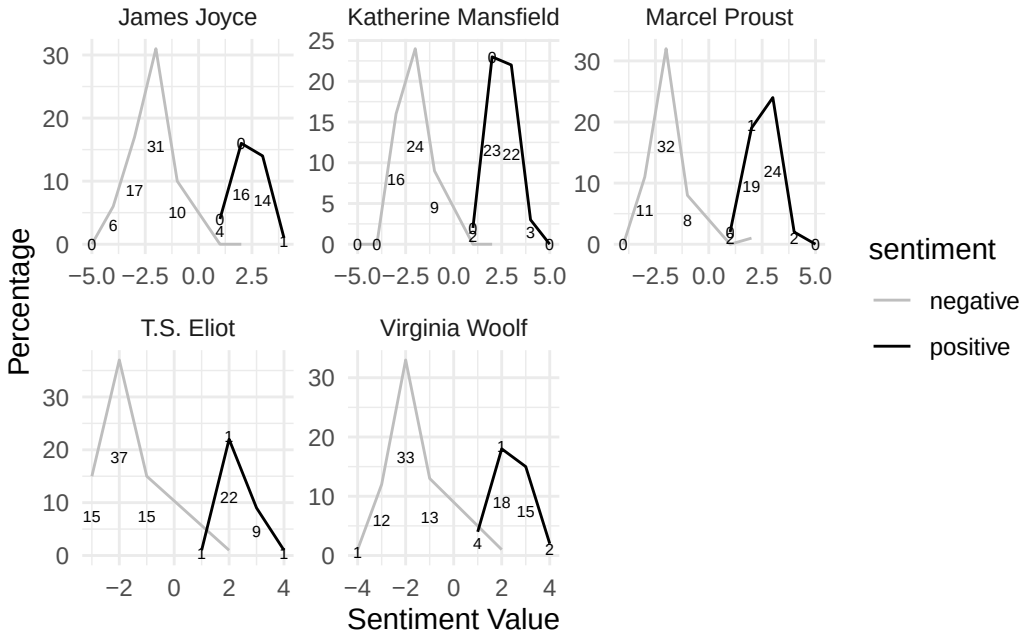
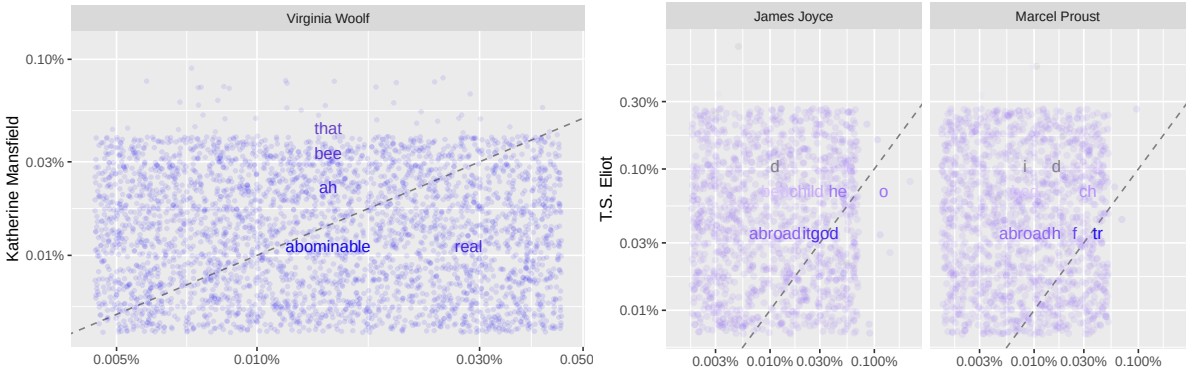



Figure 5: Numerical Sentiment Analysis of All SOC Novel Texts by Authors (AFINN lexicon)




(a) Female Authors

(b) Male Authors

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

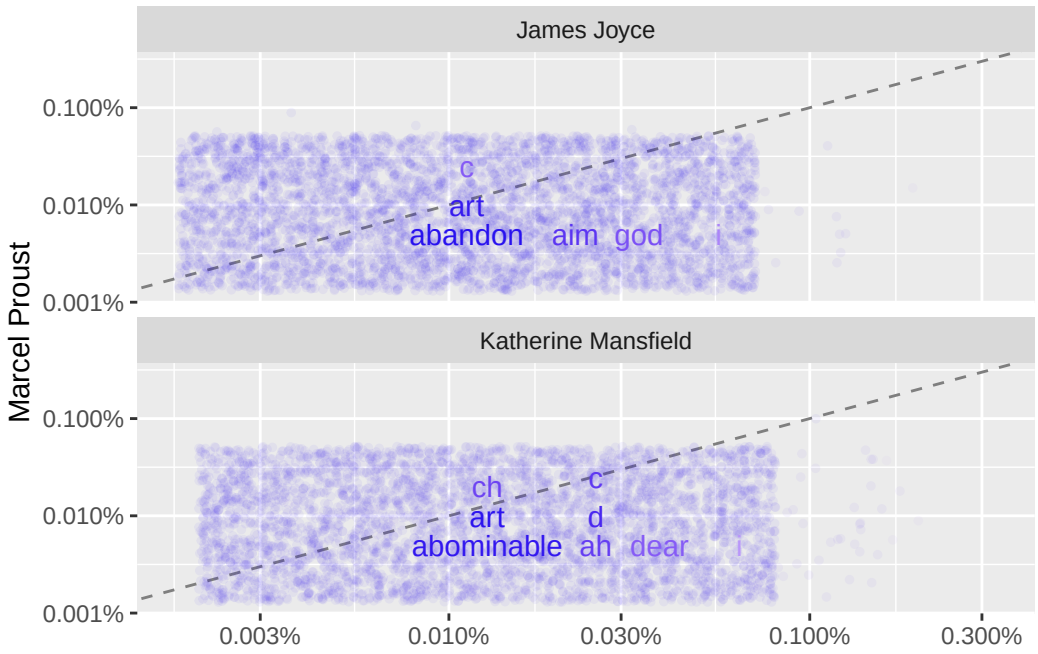



Figure 7: Comparative Analysis of Word Frequency in Transnational Stream of Consciousness Authors

### 3.3 Gendered Mental Landscape of Stream of Consciousness Novels

#### 3.3.1 Comparing SOC Literature's Female Authors

#### 3.3.2 Comparing SOC Literature's Male Authors

### 3.4 Transnational SOC Novels and Mental Health Themes

## 4 Discussion

### 4.1 Mental Health Vocabulary: Patterns and Trends

Discuss vocabulary patterns and word trends

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

pathological consequences of the constant shifts of identity in migrant characters. Most post-colonial studies contend that marginalization occasioned by the dominant group result in neurotic conditions in members of the minority group.

The novels' linguistic patterns reflect the socio-political landscape of the Western hemisphere, from late 19th century to the mid-20th century.

Stream of consciousness literature emerged as a significant literary technique in the early 20th century, particularly associated with the modernist movement in literature. Several factors contributed to its rise and prominence:

Psychological Realism: One of the primary motivations for the development of stream of consciousness was the desire to represent the inner workings of the human mind more accurately. Writers sought to capture the complexity of thought processes, emotions, and perceptions in a more realistic and nuanced manner. Response to Realism and Naturalism: Stream of consciousness literature can be seen as a reaction against the more structured and external focus of realism and naturalism. Instead of portraying characters and events objectively, authors wanted to delve deeper into subjective experiences and individual consciousness. Technological and Cultural Changes: The early 20th century was a period of rapid technological advancement and social change, including the rise of psychology as a discipline. These developments influenced writers to explore new narrative techniques and modes of representation. Literary Experimentation: Modernist writers were often interested in pushing the boundaries of traditional narrative forms and experimenting with language, style, and structure. Stream of consciousness offered a way to challenge conventional storytelling techniques and engage readers in more innovative ways. Exploration of Identity and Perception: Stream of consciousness literature allowed writers to explore themes related to identity, consciousness, and perception. By immersing readers in the internal thoughts and sensations of characters, authors could interrogate questions about the nature of reality, self-awareness, and the construction of meaning. Overall, stream of consciousness literature became a prominent feature of modernist writing due to its capacity to convey the complexity of human experience and its ability to reflect the uncertainties and ambiguities of the modern world.

The rise of modernism in literature, art, and culture during the late 19th and early 20th centuries was influenced by several interconnected factors:

Industrialization and Urbanization: The rapid industrialization and urbanization of society brought about significant changes in the way people lived, worked, and interacted. This upheaval in traditional social structures and lifestyles led to a sense of dislocation and fragmentation, which modernist artists and writers sought to capture in their work. Technological Advances: Advances in technology, particularly in transportation, communication, and mass media, reshaped human experience and perception of time and space. Innovations such as the telegraph, telephone, and photography altered how people communicated, experienced distance, and understood reality. Crisis of Faith and Meaning: The late 19th and early 20th centuries witnessed a crisis of faith and meaning, as traditional religious and philosophical beliefs came into question. Developments such as Darwin's theory of evolution, Freudian psychology, and existential philosophy challenged established notions of human existence, morality, and purpose. World Wars and Political Turmoil: The turmoil of World War I and the interwar period brought about profound changes in society, politics, and culture. The devastation of war, the collapse of empires, and the emergence of new ideologies (such as communism and

fascism) shattered old certainties and created an atmosphere of disillusionment and uncertainty. Intellectual and Cultural Exchange: The modernist movement was characterized by a spirit of intellectual and cultural exchange, with artists and writers drawing inspiration from diverse sources, including non-Western cultures, folk traditions, and avant-garde movements. This openness to new ideas and influences fostered experimentation and innovation in the arts. Quest for Individuality and Authenticity: Modernism rejected the conventions and constraints of traditional artistic forms and sought to express individual subjectivity and experience in new and unconventional ways. Artists and writers embraced themes of alienation, identity, and the search for authenticity in an increasingly fragmented and alienating world.

## 4.3 Schizophrenic and Disassociative Tendencies in Female Stream of Consciousness

## 4.4 Weaknesses

### 4.4.1 Lack of Thorough Word Cleaning

This includes words such as chapter titles, Roman numbers and personal names, affecting the integrity of data analysis ("Swann" being the most common word)

### 4.4.2 Decontextualized Literature Works and Limiting Publication Editions

Words are singled out and analyzed without context which could have affected their intended meanings, especially in such a complex genre such as stream of consciousness. The limiting novel editions also doesn't make sure that their literary integrity are maintained and the analysis might have missed important texts of other editions.

While these datasets aim to illuminate on the general time trends of childlessness with each variable measured, they are often de-contextualized for the same purpose. For instance, in Table 6, there is no clarifying note on whether people who declared themselves as 'divorced' are divorced once or more. This can be extremely narrowing, and exclude nuances when analyzing the contexts behind childlessness.

### 4.4.3 Reliance on Three Different Indices

Since the novels are chosen based on their worldwide reception and canon, practical constraints like how all novels should be of similar length are ignored. This constraint can prevent one novel having more words than others, which can critically affect the integrity of the word frequency analysis, sentiment value and word networks where one novel dominates the others and skews the results.

In addition, these authors work are expansive and so choosing a select few to bind them to the SOC genre can be limiting as the SOC genre in itself is already an amalgamation of different literary trends. Thus, this can affect the integrity of the datasets.

## 4.5 Moving Forward and Next Steps

# 5 Appendix

## 5.1 Additional Data Details

```
##| eval: true
##| echo: false
##| message: false
##| warning: false
#combined_books
```

### 5.1.1 Data Gathering

```
##| echo: false
##| message: false
##| label: tbl-reasons-strip-search
##| tbl-cap:
```

### 5.1.2 Data Cleaning

```
##| echo: false
##| message: false
##| label: tbl-items-strip-search
##| tbl-cap:
```

## 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/.

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/widyr.

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.