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

1 Introduction **1**

2 Data **3**
2.1 Measurement 3
2.2 Source Data: Project Gutenberg 4
2.2.1 Data Cleaning and Word Tokenization 5
2.2.2 Comparative Word Frequency 7
2.2.3 Generating Word Networks 7
2.3 Data Limitations 7

3 Model **7**
3.1 Model set-up 7
3.1.1 Model justification 8

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

4 Results **8**
4.1 The Dominant Vocabulary of SOC Literature 8
4.2 Sentiment Analysis 10
4.3 Gendered Mental Landscape of Stream of Consciousness Novels 10
4.3.1 Comparing SOC Literature's Female Authors 10
4.3.2 Comparing SOC Literature's Male Authors 10
4.4 Transnational SOC Novels and Mental Health Themes 10
4.5 Combined Texts: Trends, Word Networks, Bigram and Trigram Analsis 10

5 Discussion **10**
5.1 Mental Health Vocabulary: Patterns and Trends 10
5.2 Insights into Socio-Political Landscape of the West's Modernist Era 10
5.3 Schizophrenic and Disassociative Tendencies in Female Stream of Consciousness 12
5.4 Weaknesses 12
5.4.1 Lack of Thorough Word Cleaning 12
5.4.2 Decontextualized Literature Works and Limiting Publication Editions . 12
5.4.3 Uneven Novel Length and Categorization of Authors 14
5.4.4 Project Gutenberg's Focus on the Canon 14
5.5 Moving Forward and Next Steps 14

6 Appendix **15**
6.1 Additional Data Details 15
6.1.1 Data Gathering 15
6.1.2 Data Cleaning 15
6.2 Model Details 15
6.3 Posterior predictive check 15
6.4 Diagnostics 15

References **16**

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 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), `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 decontextualized and 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 contextuality. 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 depends mainly on volunteer efforts for support, and its literature is chosen and proofread by volunteers. However, the archive doesn't take copyrighted or modern items, even with permission. To submit work to the archive, you need to get copyright, scan or harvest images of book pages, spend hours proofreading and formatting, and ensure the eBook meets Project Gutenberg's requirements, including being valid HTML, correctly spelled, and compliant with Project Gutenberg's requirement. Voluntary proofreaders also have strict guidelines that they have to comply with. 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.

In terms of format, Most new Project Gutenberg eBooks are in HTML format. Project Gutenberg will check each work's the HTML validity using the W3C's online validator. The archive also askss for a plain text version whenever possible. This is because plain text has been around for a long time, is widely accessible on all devices and ensures long-term usability. PDF, Word, and other word processor formats are not used as master formats because they are harder to convert to HTML and update ("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

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.

As mentioned, to quantify the word frequency of each literary text, I first need to tokenize or separate the texts into individual chunks of words using `count()` and `unnest_tokens()` (Julia Silge and Robinson 2016).

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

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

Book ID	Text	Book	Author
2817	III At that hour when all things have repose,	Chamber Music	James Joyce
2817	O lonely watcher of the skies,	Chamber Music	James Joyce
2817	NA	Chamber Music	James Joyce

Table 2: An Examplary Table Containing Tokenzied Novel 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

Table 3: An Examplary Table Containing Word Count of Each Word within Novel Texts (James Joyce)

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

2.2.2 Comparative Word Frequency

Table 4: 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

2.2.3 Generating Word Networks

2.3 Data Limitations

3 Model

The goal of our modelling strategy is twofold. Firstly,...

Here we briefly describe the Bayesian analysis model used to investigate... Background details and diagnostics are included in Appendix 6.2.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of (**rstanarm?**). We use the default priors from `rstanarm`.

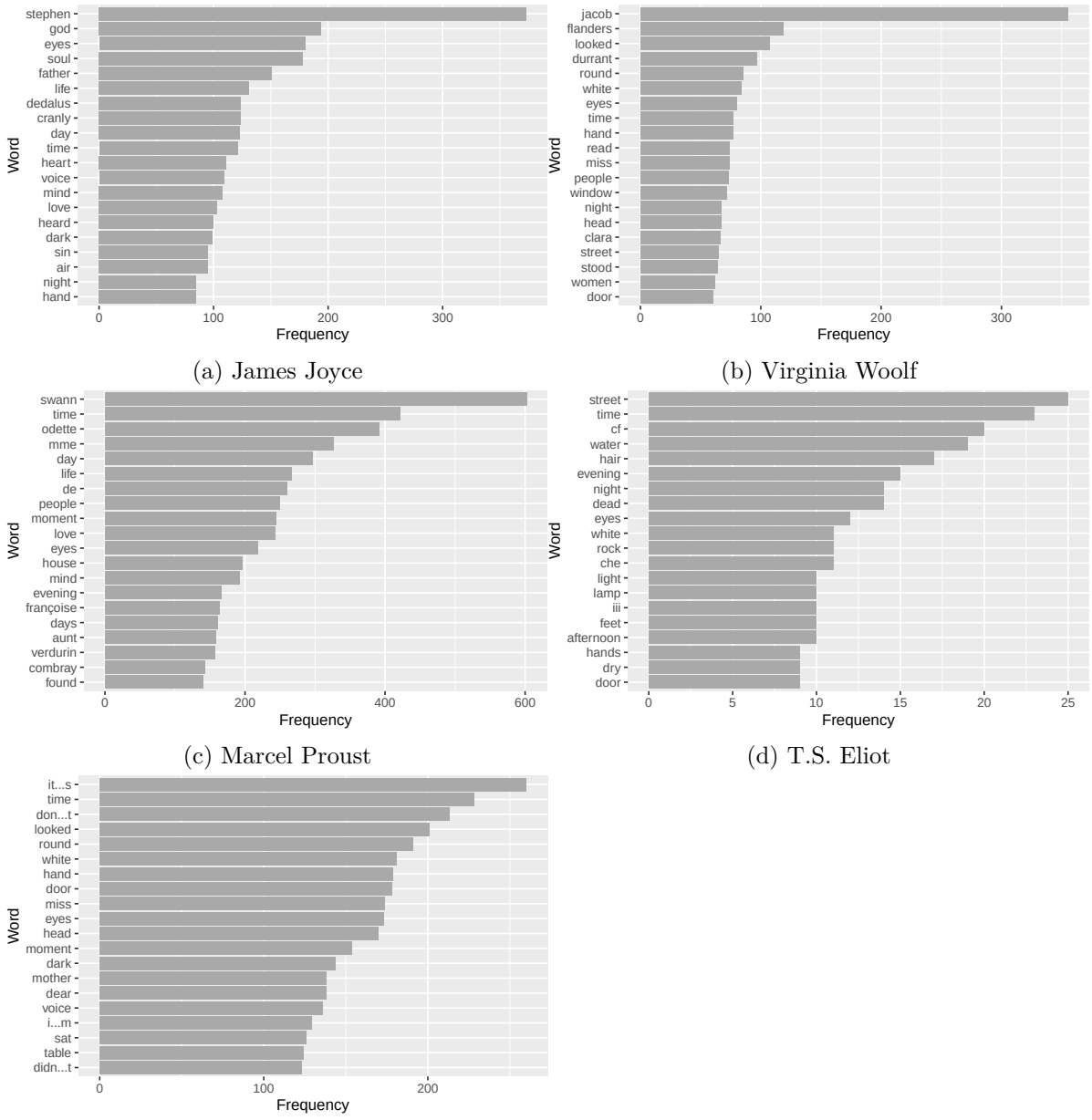
3.1.1 Model justification

We expect a negative relationship between average household income and the number of children per child care space by ward. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

4.1 The Dominant Vocabulary of SOC Literature



(a) James Joyce

(b) Virginia Woolf

(c) Marcel Proust

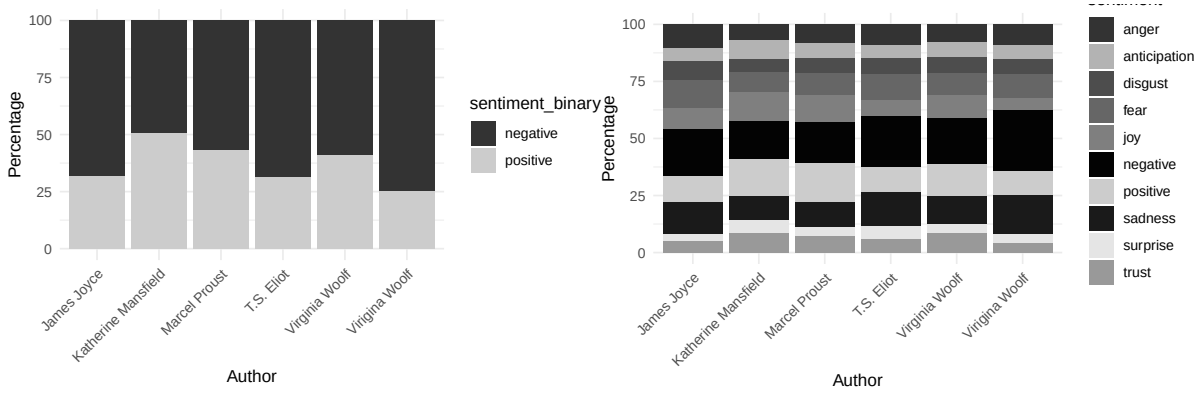
(d) T.S. Eliot

(e) Katherine Mansfield

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

4.2 Sentiment Analysis

Leveraging setiment analysis from Mohammad and Turney (2013)



(a) Bing lexicon

(b) NRC lexicon

Figure 2: Categorical Sentiment Analysis of All SOC Novel Texts by Authors

4.3 Gendered Mental Landscape of Stream of Consciousness Novels

4.3.1 Comparing SOC Literature's Female Authors

4.3.2 Comparing SOC Literature's Male Authors

4.4 Transnational SOC Novels and Mental Health Themes

4.5 Combined Texts: Trends, Word Networks, Bigram and Trigram Analsis

5 Discussion

5.1 Mental Health Vocabulary: Patterns and Trends

Discuss vocabulary patterns and word trends

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

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

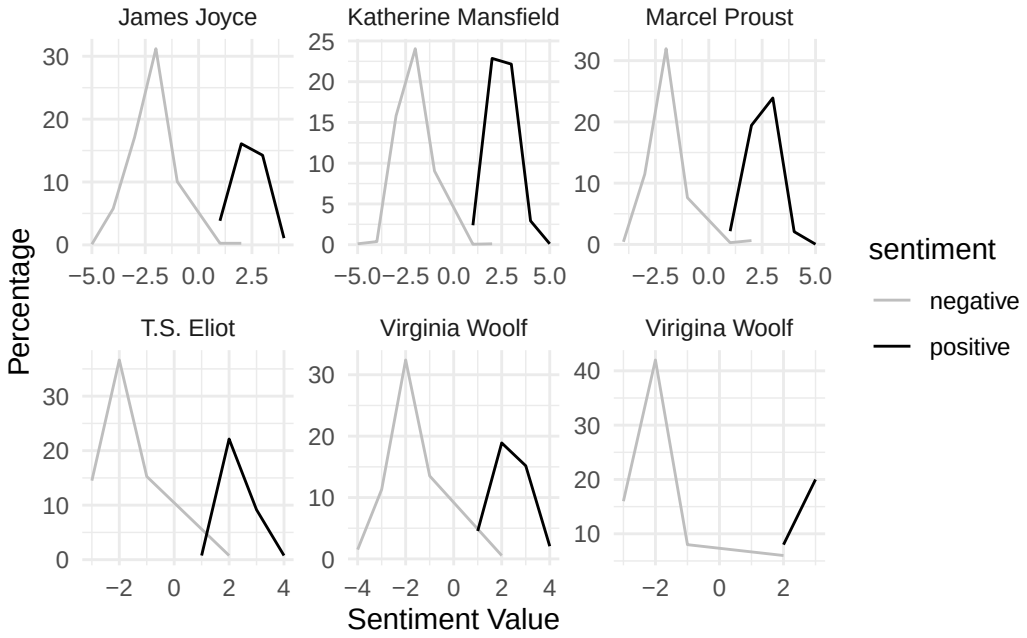
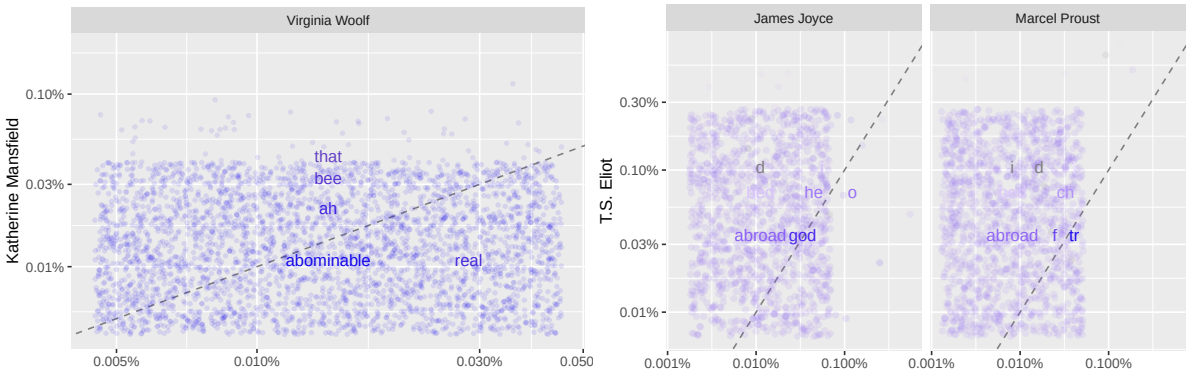


Figure 3: Numerical Sentiment Analysis of All SOC Novel Texts by Authors (AFINN lexicon)



(a) Female Authors

(b) Male Authors

Figure 4: Comparative Analysis of Word Frequency in Female and Male Stream of Consciousness Authors

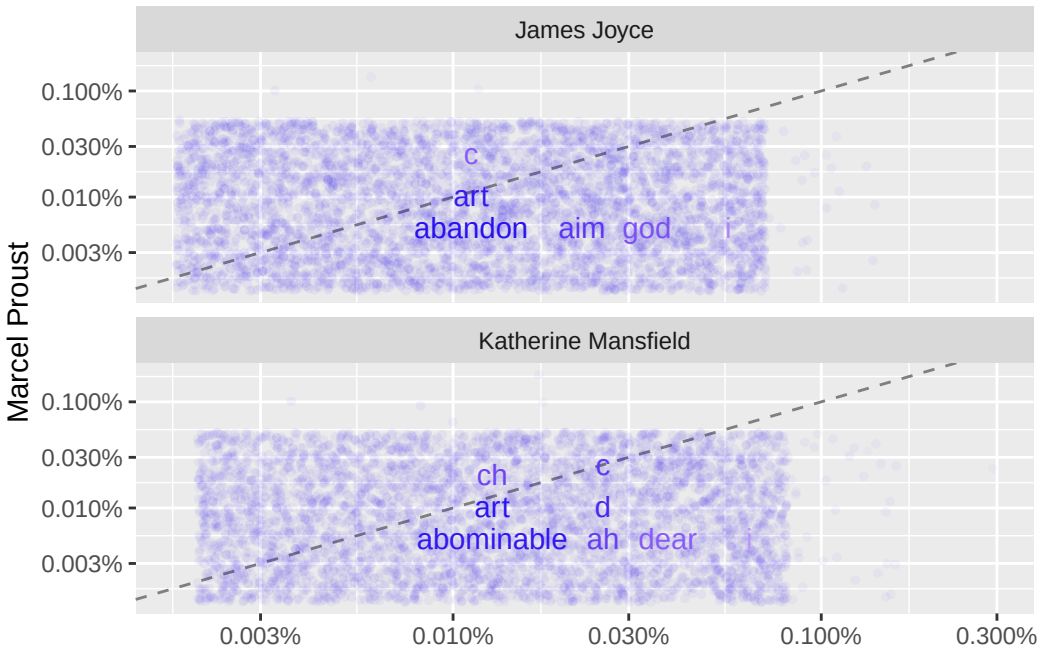


Figure 5: Comparative Analysis of Word Frequency in Transnational Stream of Consciousness Authors

5.3 Schizophrenic and Disassociative Tendencies in Female Stream of Consciousness

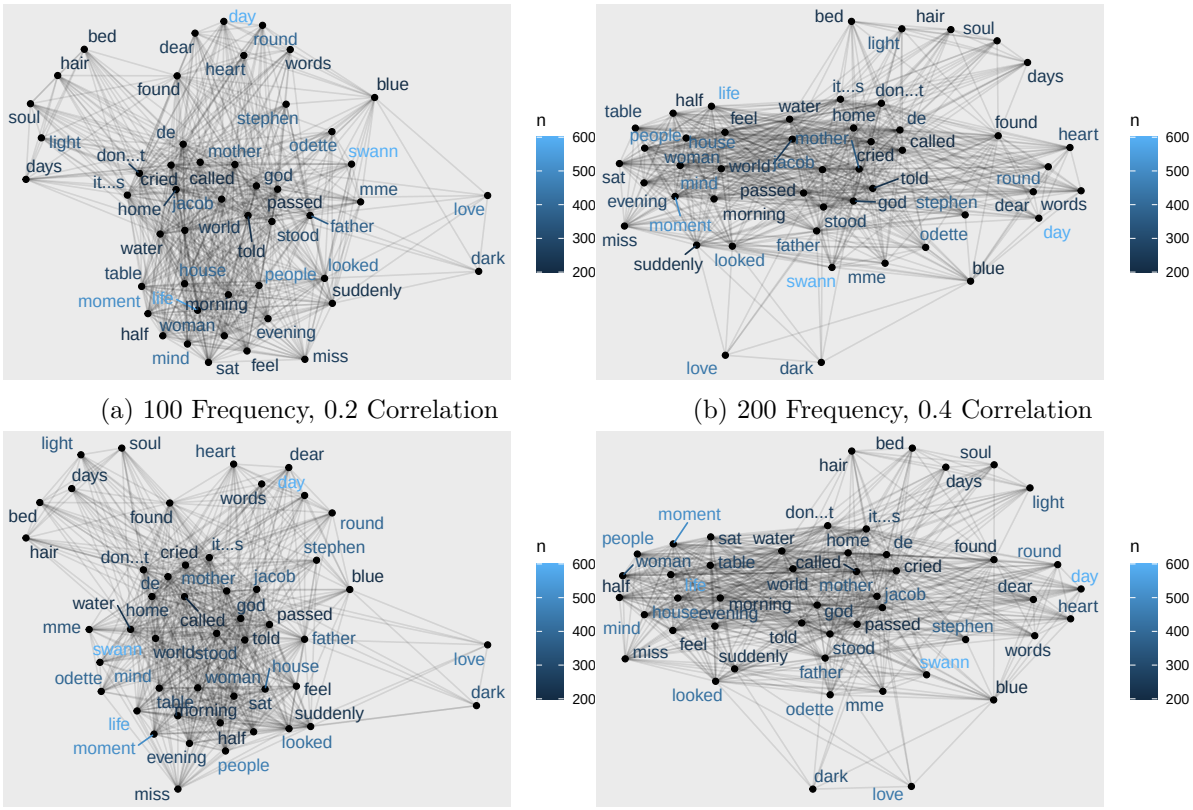
5.4 Weaknesses

5.4.1 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)

5.4.2 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.



(a) 100 Frequency, 0.2 Correlation

(b) 200 Frequency, 0.4 Correlation

(c) 200 Frequency, 0.8 Correlation

(d) 400 Frequency, 1 Correlation

Figure 6: Word Networks Measured by Frequency and Correlation when Combining All Stream of Consciousness Novels

5.4.3 Uneven Novel Length and Categorization of Authors

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.

5.4.4 Project Gutenberg's Focus on the Canon

Along the search for other stream of consciousness authors to include in the generated datasets, only a few-those that are in the SOC canon, mostly white authors-are present in Project Gutenberg, without including lesser known SOC authors, which could have steared the data analysis, still, towards their preconception: being rueful and angst-ridden

5.5 Moving Forward and Next Steps

6 Appendix

6.1 Additional Data Details

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

6.1.1 Data Gathering

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

6.1.2 Data Cleaning

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

6.2 Model Details

6.3 Posterior predictive check

6.4 Diagnostics

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/.

———. 2023. *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.