The Joyceless Landscape of Stream of Consciousness Literature*

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 4, 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 of these authors' nine novels, I aim to uncover shared linguistic patterns and gain insights into the authors' mental states. Through the analysis on SOC literature, this paper thus attempts to offer insights into themes of self-identity, anxiety, disassociation and existential contemplation within Western society's from late 19th to mid-20th century. (add one sentence on main results)

Table of contents

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

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

era of literature, spanning from late 19th century to mid-20th century ("Project Gutenberg," n.d.). (more stats and data mentioned here)

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, my estimand is the correlation between mental health-related words in SOC literature, their frequency and sentimental value as provided by Mohammad and Turney (2013). 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 2020), `igraph` (J. Silge and Robinson 2006), `widyr` (J. Silge and Robinson 2022), `ggraph` (Pedersen 2024), `textdata` (Hvitfeldt 2022), `tm` (Feinerer, Hornik, and Meyer 2008) and `knitr` (Xie 2014).