Indian Institute of Technology (Indian School of Mines) Dhanbad

Department Of Humanities And Social Sciences

NHSC507: Text Mining

Project Report

Name: Keshav Sawarn

Admission No: 24DR0340

Introduction

According to the World Population Prospects published by the United Nations, the current life expectancy in India has steeped to 70.42 years from 35.21 years in 1950. The reason behind such an increase is an improvement in healthcare facilities and a qualitatively better lifestyle. This facility is often serviced via non-governmental organisations (NGOs), government institutions, and other private organisations, many of which are supported by the government. For this, one of the most common methods has been providing the availability of health facilities that give free health treatments to the needy (Shukla, 2020).

Furthermore, when we look at human rights, we see that all of us possess human rights irrespective of race, caste, gender, sex, or place of birth. The healthcare industry is working hard to implement the rights enshrined in Article 21 of the Indian Constitution, which ensures the right to live with human dignity, and physical and psychological health becomes crucial. The industry ensures that people across all socio-economic groups and rungs can access these fundamental rights– caste, class, gender, ethnicity and so on.

Moreover, when we look at the burden on government medical facilities, which is much more than that on private facilities because, in any case, which just deteriorates, a private hospital to keep their reputation clean, refers the patient to a government hospital, thus the maximum input the government hospital has is of very sensitive cases, that's why the death ratio in government hospital is always more. Also, the caretakers of those patients put a lot of pressure on the nurses, junior doctors and senior doctors, which very often turns into verbal exchange and physical abuse within the hospital premises, because of that, sometimes the doctors beat up the patients and vice versa. When we talk about structural problems, the disproportion gap between the population and affordable medical facilities provided by government hospitals is very huge, which we can see in the doctor-patient ratio in Bihar, i.e., one doctor for 3500 people compared to the national ratio of one doctor per 1700 people (Bihar NIHFW).

The highly unaffordable cost of study, very rigorous course structure of medical, and very few medical seats due to few medical colleges in Bihar make the situation worse and lead to this mis-proportionate difference between the number of patients and the number of doctors, which is the prime reason behind the structural violence in the healthcare industry (Gupta, 2012). It has been understood that there is a massive mismatch between the number of doctors and patients, which has led to huge pressure. Now, this pressure works in two ways. Firstly, it increases the waiting time of patients, which ultimately benefits private hospitals. Secondly, it led to massive pressure on doctors and caused medical negligence where a person's human rights, right to life, and right to health, all get violated. For example, in Khagaria, Bihar, 24 rural women who had undergone tubectomy at state-run public health institutions were allegedly forced to undergo the surgical operation to prevent pregnancy without anaesthetic, leaving them awake and screaming in pain on the operating table. Kumari Pratima, one of the patients who went through the surgery, shared that-

"As I screamed in pain, four people held my hands and feet tightly as the doctor completed the job. I was administered something that left me numb only after the surgery" (Chaurasia, 2022).

Furthermore, the communication pattern between various healthcare workers, such as doctors and nurses, and their patients is reflective of the efficiency of the healthcare provided in society. The relationship between doctors and patients has evolved from a paternalistic one in earlier times (Hippocrates being the yardstick of that time) to a collaborative one in modern, contemporary society. In light of these shifts in the dynamics of the connection between physicians, nurses, and patients, there is another shift in this relationship: violence. Assaults on physicians and nurses by patients' relatives and family members have been common, both before and after the pandemic (Sarkar, 2021).

Apart from this, Paul Farmer (2003), in Pathologies of Power: Health, Human Rights, and the New War on the Poor, argued that the reason behind poor health is not just biological illness, but also social and economic, such as poverty, inequality, and social injustice, which contributes in the poor health of a person. Now, when we look at the case of the healthcare industry of Bihar, there are multiple cases of medical negligence and violence in physical, sexual, and verbal forms from both sides. It also depicts that healthcare workers are not immune to violent encounters (Kumar, 2020). Workplace violence is a major safety and health concern. Because they are at the front lines of the healthcare system and have the most direct contact with patients, violence against healthcare workers is regarded as a global health community concern.

Now, the inadequacy of healthcare resource distribution has been discovered to shape the habits of doctors, nurses, and patients. This deficiency laid the groundwork for structural violence. So, we can see that in the health care premise, structural violence, caused due to inequality, particularly in the distribution of power and resources, results in an individual's propensity for personal violence, as the people who are the victim of structural violence result into the physical violence (Galtung, 1969). It has been figured out that the healthcare industry is surrounded by both physical as well as structural violence, either at the personal level of individuals' health or at the structural level of medical negligence and inequality.

Research Objective:
To identify and analyse the dominant thematic clusters and lexical patterns in a corpus of texts relevant to public health and structural violence using text mining techniques.

Research Questions:

1. What major thematic clusters emerge from the textual corpus, and how do they relate to public health and structural violence?
2. Which terms and ideas appear most frequently and are most significant across the texts?
3. How do different grouping patterns within the corpus reflect underlying thematic or conceptual distinctions?

Methodology

This study employs a computational text analysis approach to examine dominant themes and conceptual patterns across a curated corpus of twelve books relevant to public health and structural violence. The corpus comprises digitised texts from 12 scholarly books, selected for their scholarly relevance and thematic alignment with the broader research question. These texts span theoretical, empirical, and policy-oriented literature, offering a rich source for analysing discursive formations around health, inequality, and marginalisation. The following are the 12 books used for the corpus formation:

1. Arthur Kleinman - Writing at the Margin: Discourse Between Anthropology and Medicine- University of California Press (1997)
2. Paul Farmer - Infections and Inequalities: The Modern Plagues, Updated with a New Preface- University of California Press (2001)
3. Kleinman, Arthur - The Illness Narratives: Suffering, Healing, and the Human Condition- Hachette Book Group USA (2020)
4. Marian Osterweis, Arthur Kleinman - Pain and Disability: Clinical, Behavioural, and Public Policy Perspectives, Academy Press (1987)
5. Arthur Kleinman, Matthew Basilico, Jim Yon Kim, Paul Farmer - Reimagining global health: an introduction- University of California Press (2013)
6. Iain Wilkinson, Arthur Kleinman - A Passion for Society: How We Think about Human Suffering- University of California Press (2016)
7. Arthur Kleinman, Veena Das, Margaret Lock (eds.) - Social Suffering- University of California Press (1997)
8. Alicia Ely Yamin and Paul Farmer - Power, Suffering, and the Struggle for Dignity: Human Rights Frameworks for Health and Why They Matter- University of Pennsylvania
9. Paul Farmer - Partner to the Poor, A Paul Farmer Reader (California Series in Public Anthropology) (2010)
10. Paul Farmer, Jonathan Weigel, Bill Clinton - To Repair the World, Paul Farmer Speaks to the Next Generation- University of California Press (2013)
11. Paul Farmer, Salmaan Keshavjee- Blind spot: How Neoliberalism Infiltrated Global Health- University of California Press (2014)
12. Paul Farmer - Pathologies of Power: Health, Human Rights, and the New War on the Poor

Text Preprocessing

The PDFs were transformed into plain text using the PyMuPDF library, which keeps text semantic integrity so it can be processed downstream. The extracted raw texts underwent a lemmatising process through a pipeline implemented in spaCy and filtering stopwords using NLTK. In addition to standard stopwords, a custom stopword list that excluded high-frequency pronouns and auxiliaries was also used. Numerical tokens, such as punctuation marks and special characters, were removed. All words were also made lowercase for consistency. By preprocessing the data in this manner, the analysis could focus on units of language with semantic meaning.

Lexical Pattern Analysis

Frequency-inverse document frequency (TF-IDF) weighting was used to examine prominent terms throughout the corpus. TF-IDF allows for the discovery of not only frequently used words, but also words that are distinctive across documents, thus allowing for the discovery of terms with pre-eminent local meaning. The top twenty average TF-IDF scores were displayed using bar plots to show clear lexical patterns. These bar plots helped provide an initial tour of the conceptual density and diversity in the corpus.

Thematic Clustering

The research also explored the content level of thematic clustering by transforming the "cleaned" texts into numerical vectors representing lists of term-frequency/inverse-document-frequency (TF-IDF), and then applying principal component analysis (PCA) to reduce dimensionality. The vectors were used as input to clustering algorithms to cluster documents based on semantic proximity. Two clustering approaches, partition-based or hierarchical, were used to consider different structural representations of the data. This was represented in a series of dendrograms and scatter plots of PCA-reduced components, labelled with document names to track thematic propinquities and distances. These clusters were then interpreted by many of the components' documents and the main lexical patterns to discover potential connections to a larger public health and structural violence discourse.

Topic Modeling

Simultaneously, latent topic structures were extracted using Latent Dirichlet Allocation (LDA). The algorithm was run on the already cleaned corpus and allowed for the identification of recurring thematic bundles that crossed individual documents. Topics were interpreted by viewing the top keywords for each, and subsequently using a qualitative assessment of the document space using the highest frequency of topics. The topic space was visualised using pyLDAvis, which gave the ability to visually compare the related aspects of the topics, such as their coherence and prevalence, and the distance between topics.

Analytical Orientation

This study adopts a critical computational approach, combining natural language processing and reflexive sociological logic. Although the tools are algorithmically designed (i.e., based on computer programming), I still maintain the process of theme identification, clustering, and topic significance conceptually around the theoretical aspects of sociologies of health and structural violence. Therefore, the method is not geared towards predictive modelling, but rather aims to discover latent structures of meaning and relationalities found within the text, and allow for a discussion surrounding how discourses on health are constructed, contested, and reproduced in the literature.

Result

Figure 1 portrays the cosine similarity heatmap, which reveals patterns of thematic and conceptual overlap among twelve foundational texts in medical anthropology and global health. A prominent cluster emerges around the works of Paul Farmer-Pathologies of Power, Partner to the Poor, To Repair the World, and Blind Spot- which exhibit high mutual similarity (cosine values above 0.8), indicating consistent conceptual terrain across his writings, particularly around structural violence, social suffering, and global health equity. Similarly, Arthur Kleinman's key works- The Illness Narratives, Writing at the Margin, and the co-edited volume Social Suffering- display moderate to high similarity scores (0.6–0.7), reflecting shared concerns with subjectivity, clinical experience, and the moral dimensions of care.

Interestingly, the volume Social Suffering and Reimagining Global Health (co-authored by Farmer and Kleinman) occupies a central position, with relatively high similarity to multiple texts, suggesting their function as conceptual bridges synthesising themes across authors and disciplines. In contrast, Alicia Ely Yamin's Power, Suffering, and the Struggle for Dignity and Pain and Disability by Osterweis et al. appear as thematic outliers, with lower similarity scores (often below 0.4), likely due to their distinct legal-political and public health policy orientations, respectively. Overall, the heatmap reveals two dense clusters- one centred on Farmer and another on Kleinman- linked through shared discourses of suffering, health justice, and the critique of biomedical reductionism, while also highlighting texts that contribute divergent but complementary perspectives.

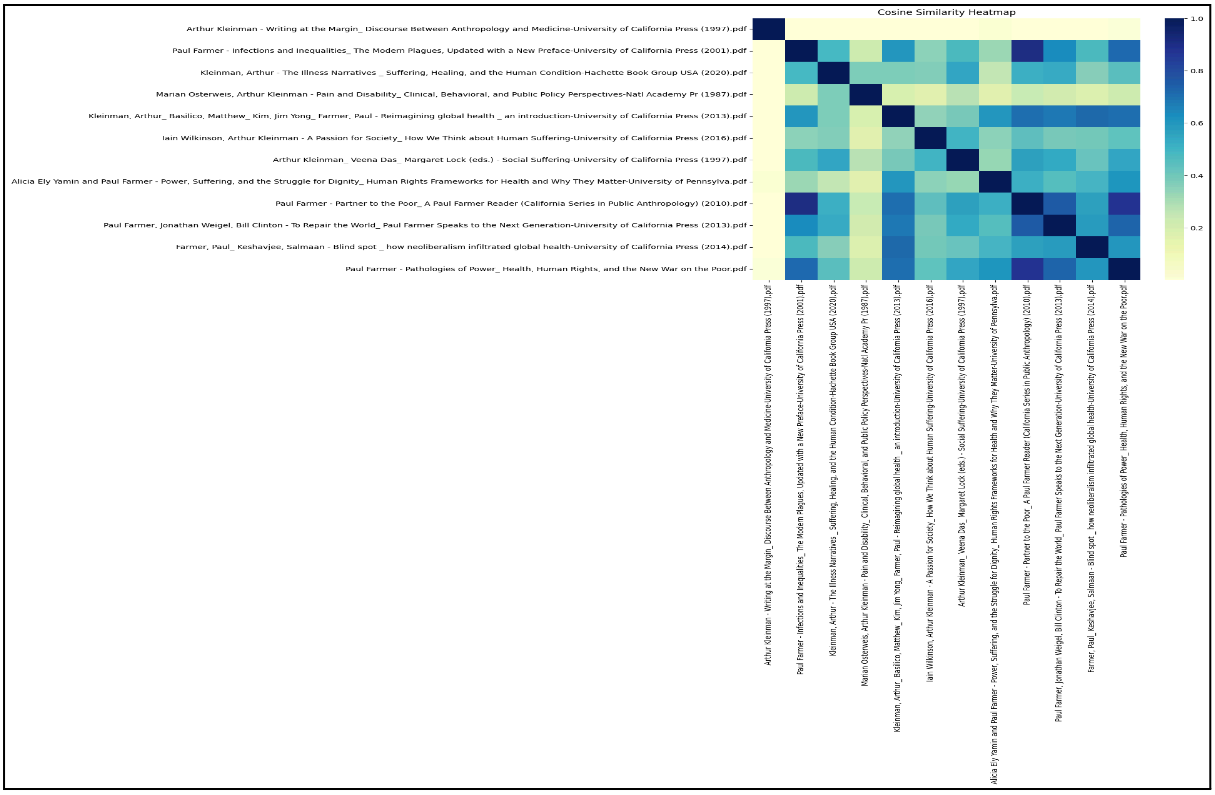

Figure 1

Figure 2 presents the PCA plot, which visualises the distribution of twelve texts based on their TF-IDF scores, reduced to two principal components and grouped into three clusters using K-Means. Cluster 1 (blue) primarily contains texts by Paul Farmer and Arthur Kleinman that focus on structural violence, suffering, and ethical responses in healthcare, suggesting a strong thematic coherence in their critique of biomedicine and advocacy for health justice. Cluster 2 (orange) encompasses more dispersed texts, including those by Marian Osterweis and Alicia Ely Yamin, which deviate somewhat from the Farmer-Kleinman core by emphasising either public health systems or rights-based legal frameworks, indicating broader or more policy-oriented concerns. Interestingly, Arthur Kleinman's The Illness Narratives- Cluster 3 (green)- is an outlier, spatially distant from other works, suggesting a distinct vocabulary or conceptual structure. This may reflect its strong focus on clinical ethnography and patient narratives, which differ lexically and thematically from more overtly political or global health texts. The clustering supports the earlier cosine similarity findings, showing a central thematic axis around Farmer and Kleinman, with variations introduced by works on disability, rights, or public health policy.

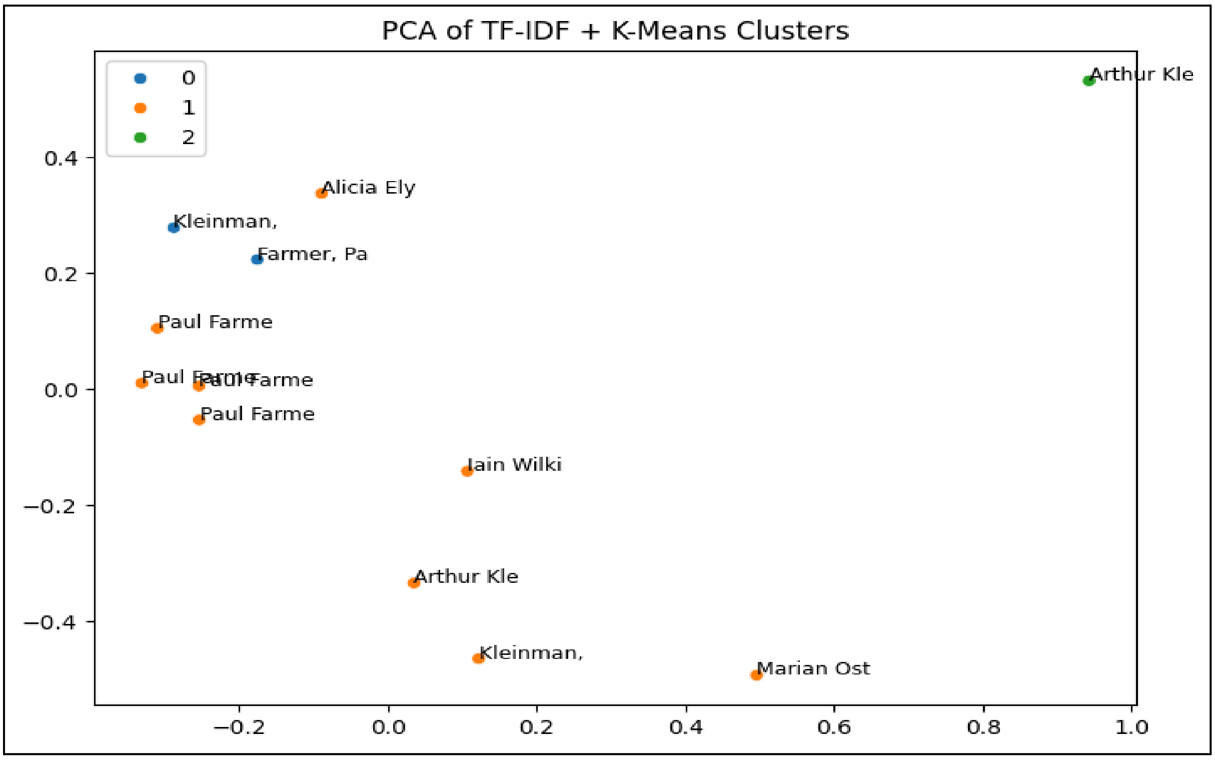

Figure 2

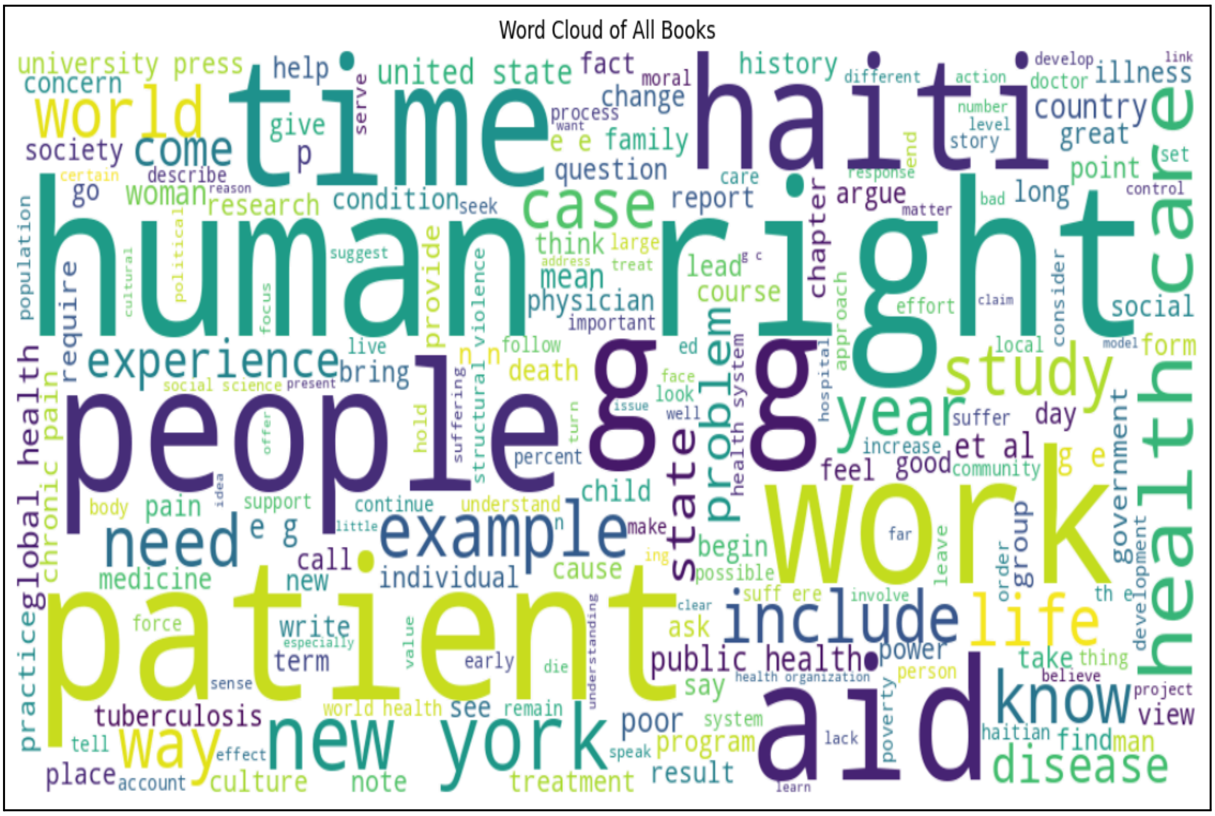

Figure 3

Figure 3 represents the word cloud, which brings together the highest frequency words from all twelve books so that recurrent themes and conceptual preoccupations emerge. Words like "patient", "people", "human", "rights", and "healthcare" specifically communicate sustained investments in person-centred care, human rights, and systemic injustice in terms of healthcare systems. The words "Haiti", "aid", "tuberculosis", and "structural violence" demonstrate a geographic and an epidemiological grounding of the texts, and it seems particularly salient in the case of Paul Farmer. The presence of words like "experience", "pain", "suffering", and "culture" accentuates phenomenological and ethnographical approaches associated with Kleinman's work. The identified words "New York", "United States", and "global health" accentuate a transnational and comparative analytical framework. Collectively, the lexical field reveals an epistemic disposition that connects the clinical telling to larger structures of inequality and highlights how biopolitical concerns reconcile with ethnographical ethics and health justice engagement.

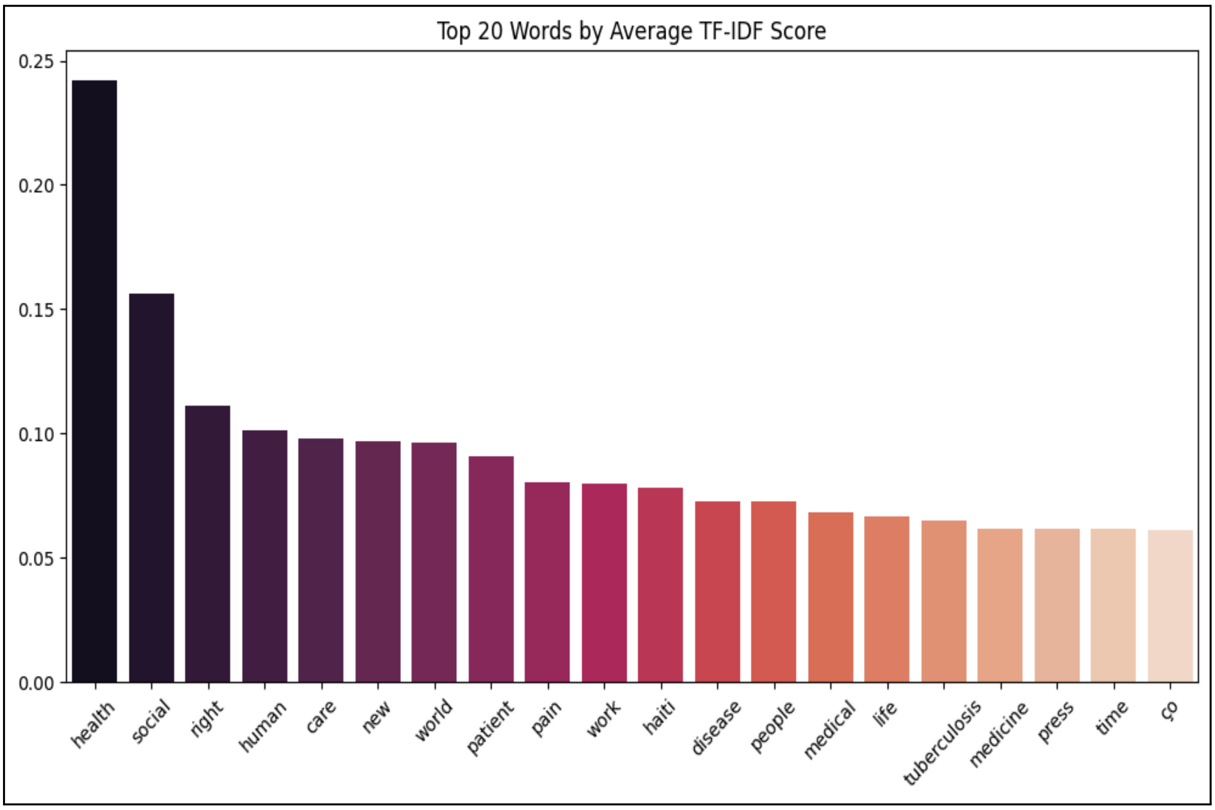

Figure 4

The concept of the "Top 20 Words by Average TF-IDF Score" is represented in **Figure 4**, which is a bar chart which allows a visualisation of the most distinctive words in a set of documents, based on their TF-IDF (Term Frequency–Inverse Document Frequency) scores. TF-IDF is a widely accepted text mining technique used to determine the importance of a word in the context of a single document compared to a collection (corpus) of documents. In this chart, "health" is the highest word, suggesting its strong significance and distinctiveness within the texts making up the corpus. In addition, other words with high scores like "social," "right," "human," and "care" also point toward a thematic focus of social justice, human rights, and healthcare. Some other words like "patient," "pain," "disease," "medical," "medicine," and "tuberculosis" suggest a medical or health-focused issue, potentially related to humanitarian or public health dialogues. The word "Haiti" might refer to a geographical or case-study reference in this corpus with potential health crisis implications or interventions. Overall, the chart indicates a strong convergence around healthcare, human rights, and social welfare. Together, these themes seem to represent the main, documented subject matter of the source documents.

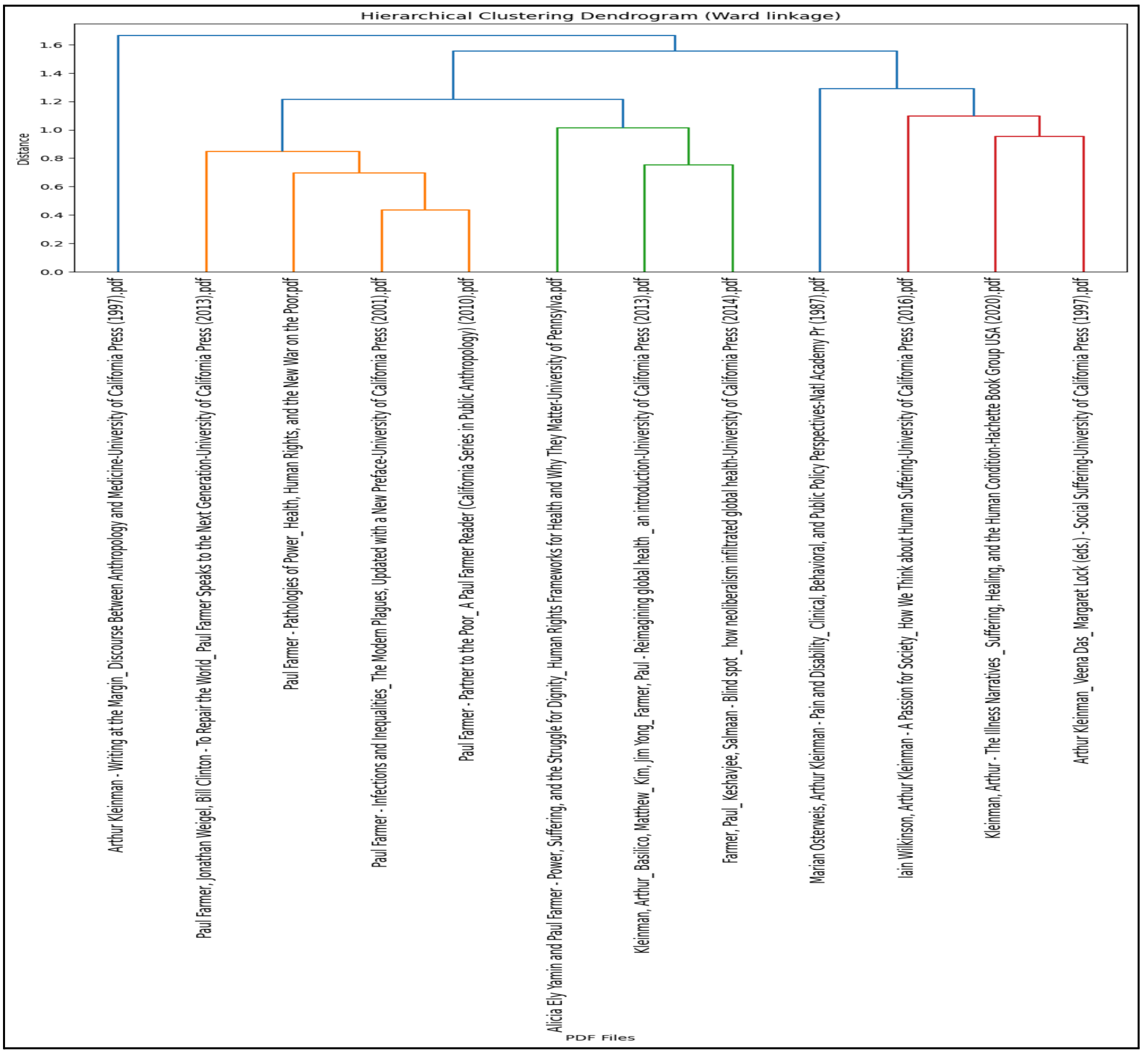

Figure 5

Figure 5 presents the dendrogram, which is labelled "Hierarchical Clustering Dendrogram (Ward linkage)" and represents the similarity relationships between a set of documents, most of which are authored or co-authored by well-known medical anthropologists such as Paul Farmer and Arthur Kleinman. Ward's method (minimising total within-cluster variance) grouped the documents by some lexical features extracted via TF-IDF or some other vectorisation procedure. In the vertical axis labelled "Distance", dissimilarity between clusters can be seen: the greater the distance from the joining point between each document or cluster, the greater the dissimilarity between them. Within this dendrogram, we can see several clusters of Paul Farmer's writing, who appears several times throughout the analysis. Farmer's writings cluster closely, suggesting that they share thematic and lexical coherence, and that this similarity particularly accords with his writings on matters of health, inequality, and global justice. Similarly, the works of Arthur Kleinman were also clustered but are separate from Farmer's cluster. This suggests that Kleinman's works on suffering, caregiving and cultural dimensions of illness also share thematic and lexical coherence. Some sub-clustered collections link collaborative writings and share some thematic concerns, such as global health policy or structural violence. The visual structure embedded in the dendrogram can show how the texts relate to one another conceptually and indicate thematic continuities, or gaps and overlaps in the literature.

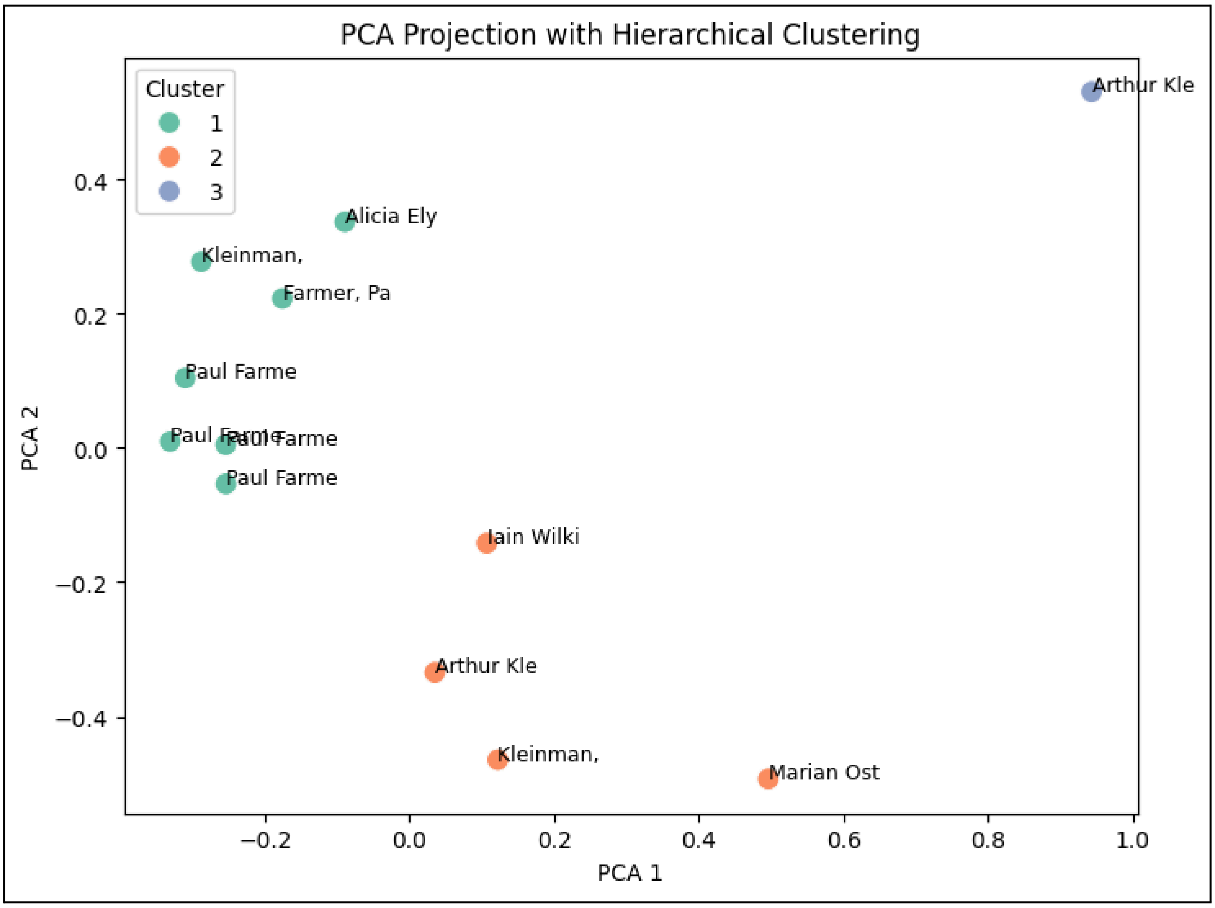

Figure 6

With hierarchical clustering, the PCA projecting plot in **Figure 6** provides a visual summary of how the texts cluster based on thematic and lexical similarities. Each document is represented as a point in a two-dimensional space created by principal component analysis, reducing the data complexity while maintaining the most explanatory elements of the original data. The colour of each point corresponds to the cluster it was observed in based on hierarchical clustering. Cluster 1 (green) consists mainly of texts by Paul Farmer and some of his collaborators (e.g., Alicia Ely Yamin), indicating a thematic focus among the texts on themes like global health, human rights, and social medicine. Cluster 2 (orange) contains texts from authors like Arthur Kleinman and Marian Osterveer. Although they include discussions related to public health, the cluster also suggests a greater emphasis on the outputs of medical anthropological work and critical perspectives on health systems. In Cluster 3 (blue), there is only an outlier representing an Arthur Kleinman text. However, technically a part of the corpus of texts clustering together, it appears in isolation in the plot, suggesting a differentiation in some conceptual or linguistic content. The spatial arrangement of the points in the plot informs us how these texts have diverged or converged concerning their focal points in the development of the text, ultimately offering a map of scholarly affinities realised in the corpus.

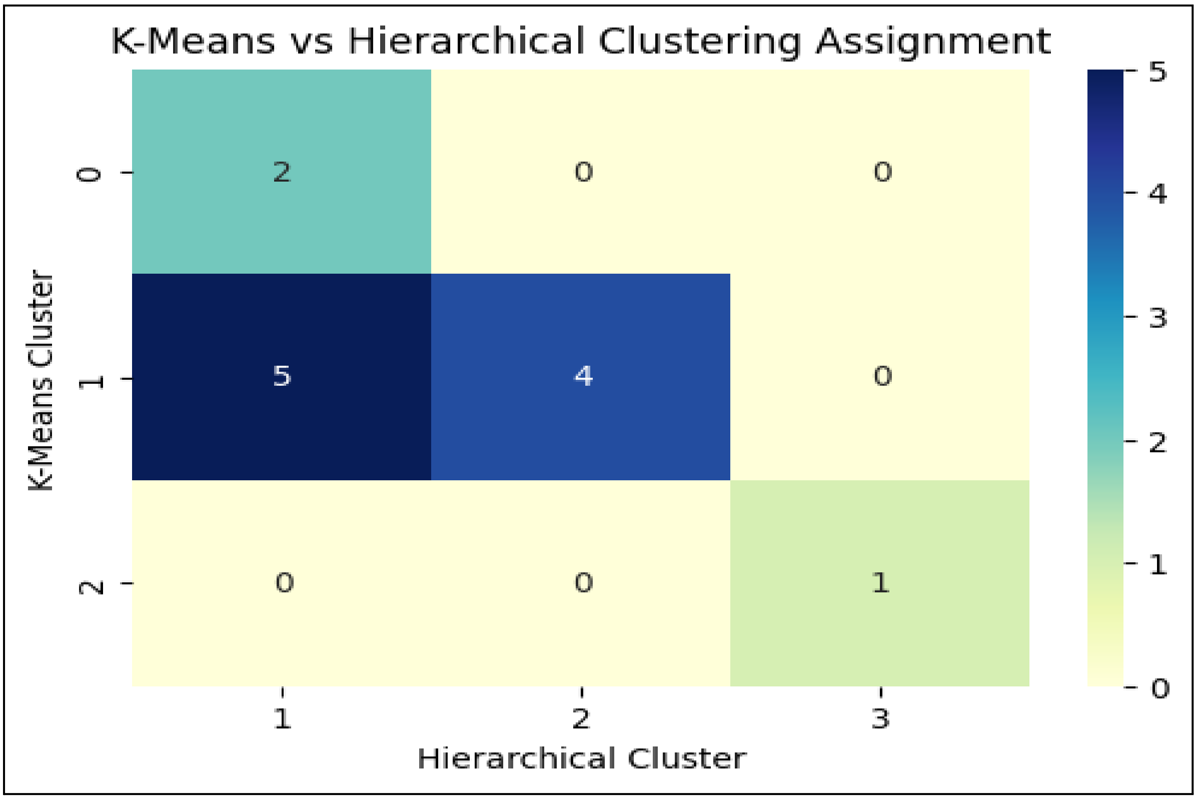

Figure 7

This heatmap in **Figure 7** shows the similarity in cluster assignments for K-means and Hierarchical Clustering, suggesting the level of concordance. The values inside the cells reflect the number of documents members of each K-Means cluster (rows) shared with each Hierarchical cluster (columns). The most considerable overlap is between K-Means Cluster 1 and Hierarchical Clusters 1 and 2, with they sharing 5 and 4 documents, respectively, meaning that the collection of texts contains inner consistency identified by both methods, even with some differences in boundaries. K-Means Cluster 0 mostly overlaps with Hierarchical Cluster 1, while K-Means Cluster 2 overlaps only with Hierarchical Cluster 3, indicating that some thematically outlying material is still recognised with some consistency across the algorithms. This matrix validates cluster results and shows where K-means and Hierarchical clustering converge and diverge, aiding in a more nuanced discussion about the similarities in the underlying texts.

Conclusion

This study used computational text mining tools to identify the thematic clusters and lexical features in a collection of twelve foundational public health texts and their intersection with structural violence. The analysis identified two primary thematic clusters, aligned mostly with the work of Paul Farmer and Arthur Kleinman, with similar themes around structural violence, health justice, and critiques of biomedical reductionism. Farmer and Kleinman's texts show substantial conceptual overlap, while other scholars, such as Alicia Ely Yamin and Marian Osterweis, inject different viewpoints, especially into legalistic frameworks and public health policy. This study identified dominant themes such as suffering, health inequity, and human rights through TF-IDF, PCA, and LDA tools. Text mining highlighted the dynamic relationship between health discourses and sociopolitical realities. The clustering and topic modelling also provided evidence of coherence and dissonance across the documents. Using text mining tools with interpretive sociology has produced important insights into how structural violence and public health can be studied together in academic texts. They provide a useful basis for future research on this topic.

References

Chaurasia, M., & *The Times of India*. (2022, November 17). *Surgery without anaesthesia in Bihar hospital: Pinned down by four people and shrieking, women operated on*. *The Times of India*. https://timesofindia.indiatimes.com/city/patna/surgery-without-anaesthesia-in-bihar-hospital-pinned-down-by-four-people-and-shrieking-women-operated-on/articleshow/95566412.cms

Galtung, J. (1969). *Violence, peace, and peace research*. *Journal of Peace Research, 6*(3), 167–191. https://doi.org/10.1177/002234336900600301

Gupta, A. (2012). *Red tape: Bureaucracy, structural violence, and poverty in India*. Orient Blackswan.

Farmer, P. (2003). *Pathologies of power: Health, human rights, and the new war on the poor*. *North American Dialogue, 6*(1), 1–4. https://doi.org/10.1525/nad.2003.6.1.1

Kumar, P. S., Betadur, D., & Chandermani. (2020). *Study on mitigation of workplace violence in hospitals*. *Medical Journal, Armed Forces India, 76*(3), 298–302. https://doi.org/10.1016/j.mjafi.2019.09.003

Sarkar, T. (2021). *Understanding co-existence of violence and care – A sociological analysis of workplace violence in government hospitals in West Bengal, India*. Knowledge UChicago. https://doi.org/10.6082/uchicago.3106

Shukla, S. (2020). Tort of medical negligence in India. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3621457