Topic Trend Analysis in VA Health Systems Research Forums

Abstract

This study applies NLP techniques to analyze publications from the VA Health Systems Research Forum, aiming to uncover trends and sentiments in veteran healthcare system discussions. By implementing LDA, LSTM, and BERT models, the analysis revealed significant insights into the progression of key community topics such as telehealth, PTSD, and caregiver. Sentiment analysis highlighted the optimism surrounding various health topic publications. These findings demonstrate the potential for tailored NLP applications to improve health information accessibility, inform policy development, and improve healthcare delivery strategies for veterans.

Keywords: Veterans Affairs, Topic Modeling, Sentiment Analysis

Introduction (Motivation for this Work)

The Veterans Affairs (VA) manages the largest integrated healthcare system in the United States, which includes 170 medical centers and 1,193 outpatient facilities (US Department of Veterans Affairs, 2024a). Each year, the VA allocates approximately $68 billion in services to more than nine million enrolled veterans. Fundamental to its mission, the VA operates a robust health systems research (HSR) component tasked with examining all facets of healthcare delivery, including patient care, cost efficiency, and level of quality. The findings from these investigations are disseminated to the public through quarterly "Forum" publications by the VA Health Systems Research division (US Department of Veterans Affairs, 2024b). This rich repository of data provides a unique opportunity for conducting a topic trend analysis, which can reveal both longstanding research themes within the VA system. These insights are important, as they inform on areas that have consistently dominated the landscape of VA research while also pinpointing new challenges and priorities that may influence future healthcare policies and practices.

This project aims to apply advanced natural language processing (NLP) techniques to systematically analyze these publications. The objective is to identify and track topic trends over time, thereby offering a clearer understanding of the research focus and the potential implications for policy-making and healthcare service delivery in the veterans' health system.

Background (Literature Review)

Military veterans are a unique population with complex health needs stemming from severe physical injuries and mental health disorders. These health challenges are compounded by various social determinants of health (SDOH) such as education level, employment status, and household income, which significantly influence a veteran's ability to access and understand health information. The concept of health literacy is particularly important since it ensures that veterans are well-informed and able to actively participate in managing their own care. Prior research has underscored a persistent gap in health information access among veterans. Taylor et al. (2020) highlighted the necessity for deeper investigation into the topics of mental health information, particularly in understanding how mental health services are utilized and supported by policies within the VA healthcare system.

Moreover, the impact of the digital divide on health inequality among veterans has been a focus of recent studies. Swed, Sheehan, and Butler (2020) explored how disparities in internet usage could affect health outcomes. They found that veterans with less frequent internet access were significantly "more likely to report poor self-rated health, emphasizing the digital divide as a barrier to health equity" (Swed et al., 2020, p.12). Innovations in NLP have provided new methodologies to address these challenges. For example, Mitra et al. (2023) utilized NLP to extract SDOH factors from unstructured clinical notes, which highlighted several SDOHs, including: social isolation, job or financial insecurity, housing instability, legal problems, barriers to care, violence, transition of care, and food insecurity.

Furthermore, the application of NLP in healthcare research has demonstrated potential beyond SDOH extraction. Studies such as those by Houston et al. (2013) show that veterans are likely to seek information on specific health issues like Alzheimer's disease more frequently than their non-VA counterparts. This points to the necessity for accessible, targeted health information that can enhance veteran health literacy and outcomes. Significant work has been performed using NLP to streamline research processes and surveillance in areas like breast cancer (Carrell et al., 2014) and cervical cancer (Oliveira et al., 2020), showcasing the versatility and impact of NLP technologies in improving healthcare delivery and research efficacy.

Problem Statement

Despite the extensive resources and care provided by the VA, there remains a significant gap in veterans' access to valuable health information. This gap affects their ability to manage personal health effectively and impacts broader health outcomes across the veteran population. While previous studies have explored aspects of health information accessibility, there is a notable scarcity of research leveraging NLP to improve information dissemination and health literacy among veterans. The existing research predominantly focuses on the application of NLP in clinical settings for patient care management and disease surveillance. However, the potential for NLP to improve the accessibility and personalization of health information for veterans, remains largely untapped. This study seeks to fill this gap by utilizing NLP to examine topics pertinent to the veteran population.

Success Measures

The success of this project was determined by its ability to systematically identify and analyze evolving trends in topics covered by the VA Health Systems Research Forum. Success was demonstrated through: (1) Frequency Trends, where significant changes in the prominence of key topics were detected over time, indicating shifts in healthcare priorities; (2) Sentiment Analysis, by providing clear insights into the sentiments (positive, neutral, negative) surrounding these topics, offering a deeper understanding of how discussions are framed within VA research; and (3) Topic Coherence, through a meaningful breakdown of related themes, which align with the broader research goals of the VA. These measures ensure that the project identifies trends within the VA's evolving healthcare priorities.

The baseline model for this project used Latent Dirichlet Allocation (LDA) to analyze text data from the VA Health Systems Research Forum publications. This model utilized basic text mining processes such as frequency analysis to track the prevalence of key health topics over time. In addition, it incorporated simple sentiment analysis tools to gauge the general sentiment associated with these health topics. The LDA model served as an initial framework to establish a foundational understanding of the topic trends and sentiment distributions within the VA health discussions, providing preliminary insight into the evolving themes of veteran health system research.

Methods

A total of 78 quarterly publications were extracted from the VA Health Service Research Forum (US Department of Veterans Affairs, 2024b). The forum is publicly available, and includes publications spanning from 1998 to 2024. All publications were downloaded and stored in a Google Drive folder, where they could be processed using Colab. For the extraction of text data, a Python script utilizing the PyPDF2 library was implemented to read and convert PDF content into text format. This allowed for the automated parsing of each document, ensuring that all textual information was accurately captured and prepared for subsequent analysis. Following text extraction, the collected data underwent a preliminary cleaning process to remove irrelevant sections such as headers, footers, and any non-textual elements like images and tables. NLP was applied to the cleaned text to identify prevalent themes and trends across the publications. These processes were facilitated using the NLTK and Gensim libraries within the Python environment, providing a robust framework for detailed textual analysis and topic modeling.

The selection of NLP techniques for this project were appropriate for addressing the problems identified in analyzing VA Health Systems Research Forum publications. NLP is fundamental for extracting meaningful patterns from large volumes of unstructured text, making it ideal for this project's focus on textual data analysis. Long Short-Term Memory (LSTM) networks were well suited for understanding context in text sequences, essential for tracking changes in topic prominence over time. The Bidirectional Encoder Representations from Transformers (BERT) model, known for its ability to comprehend the context of words by looking at surrounding text, was important for accurately analyzing sentiment and thematic coherence in health discussions. Sentiment analysis helped interpret how topics were framed, influencing policy and public perception, while frequency analysis quantified how often specific topics appeared, providing insights into evolving research priorities. Collectively, these methods addressed the complexity of the dataset.

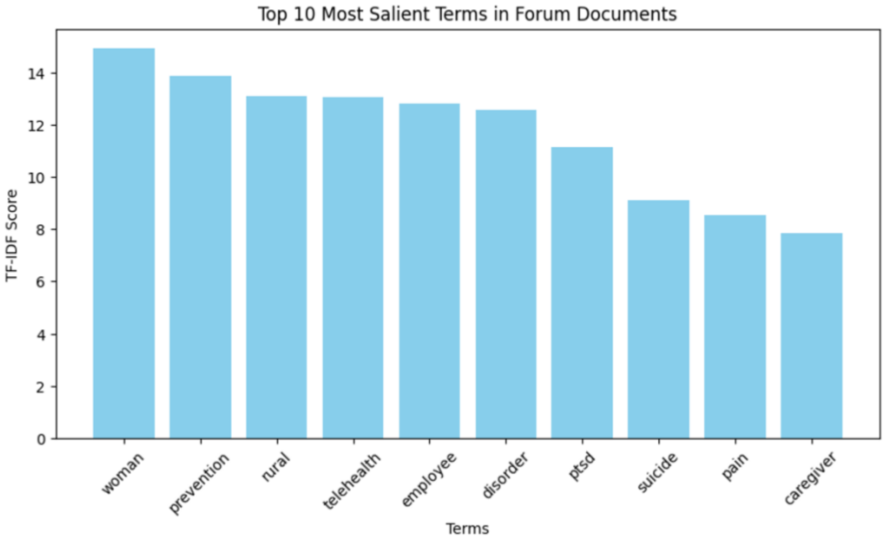
In the initial phases of the project, key lessons emerged which significantly influenced the improvement of subsequent iterations. Comprehensive data preprocessing was identified as necessary when initial model results were skewed by inconsistencies and specialized jargon, leading to the integration of text normalization and a custom dictionary. This improved the models' accuracy in interpreting the data. The early models also highlighted the challenges of model complexity and overfitting. Therefore, LSTM and BERT models were incorporated to better manage the complexities and nuances of the dataset. In addition, the initial sentiment analysis tools were inaccurately classifying sentiments due to the unique context of VA communications, prompting adjustments to the sentiment analysis algorithms to reflect linguistic nuances more accurately. Implementing a feedback loop for continuous model refinement based on intermediate findings also enhanced the project's alignment with its objectives. Scalability and processing efficiency were addressed by optimizing data handling and employing parallel processing techniques, significantly improving processing times and reducing resource usage.

During the data analysis, several unusual patterns emerged, particularly in the distribution of topics and the models' interactions with the data. For instance, there was a cyclical trend in the discussion of certain health topics such as mental health and telehealth, which seemed to peak during certain months. The LSTM and BERT models interacted with the data in ways that revealed the depth of these cyclical trends, but initially, they exhibited unexpected loss patterns during training phases, showing spikes in loss values at certain epochs which suggested sensitivity to specific features or overfitting to less relevant data segments. These anomalies prompted a deeper investigation into the model parameters and training processes, leading to adjustments such as modifying the learning rate and adding dropout layers to mitigate overfitting.

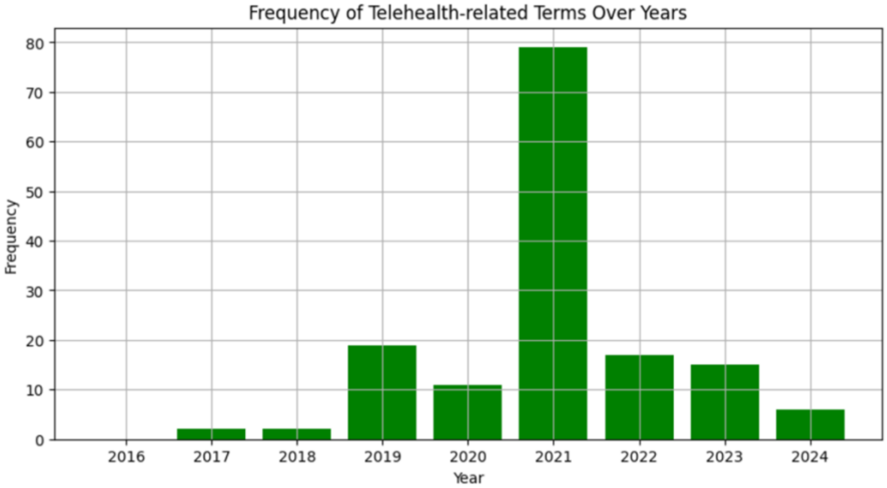
The observed patterns aligned well with findings from other researchers, particularly in the areas of topic prevalence and the challenges of sentiment analysis in healthcare-related texts. For example, cyclical trends in the discussion of telehealth and mental health were consistent with Swed, Sheehan, and Butler (2020), who identified temporal influences on health-related topics, often driven by policy changes or societal events. Similarly, the challenge of accurately interpreting sentiment in nuanced healthcare communications echoed findings by Mitra et al. (2023), who emphasized the need for domain-specific adjustments in NLP algorithms to improve performance in extracting social determinants of health. Other researchers, such as Carrell et al. (2014), addressed similar challenges by incorporating tailored preprocessing techniques and refining model architectures to better adapt to specialized datasets. Inspired by these approaches, the implementation of custom text normalization, domain-specific dictionaries, and iterative model tuning were incorporated to align the analysis more closely with the complex nature of VA healthcare communications.

Results and Discussion

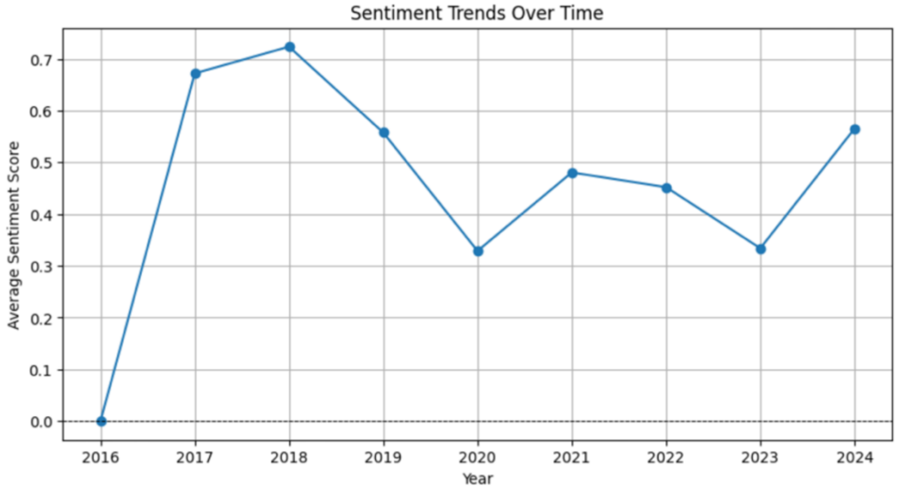
In this section, a detailed analysis is provided for the trends and sentiments identified in VA Health Systems Research Forum publications using various NLP techniques. These visual representations highlight the shifts in topic prominence and sentiment over time, offering insights into the evolving dynamics of veteran healthcare discussions. The following chart presents the top 10 terms identified in the Forum publications.



Several prominent terms were examined, and the following frequency trends analysis revealed significant temporal patterns in the prominence "telehealth". For instance, telehealth discussions surged during the COVID-19 pandemic, reflecting its increased relevance in VA healthcare delivery.



The following sentiment analysis provides valuable insights into the tone surrounding these research topics. The analysis revealed a mix of positive and neutral sentiments, reflecting optimism about treatment advancements and ongoing concerns about service gaps. Topic discussions were predominantly positive, emphasizing its potential to improve healthcare access, though occasional negative sentiments highlighted challenges such as cost and technological barriers.



The coherence score for the LDA model was computed using Gensim's CoherenceModel, assessing the semantic coherence of topics produced by the model. This metric was obtained by initializing the CoherenceModel with the LDA model, processed texts, and the dictionary linking words to integer IDs, using the 'c_v' coherence type which evaluates pairwise word similarity within a topic. The coherence score was approximately 0.3295, which serves as an indicator of topic quality. While indicative of some degree of coherence, this low score suggests that there is room for improvement in topic quality and interpretability, pointing to the need for further refinements in model parameters.

Limitations

This study includes several limitations. First, the reliance on publications from a single source, may introduce a selection bias, limiting the generalizability of the findings to other veteran healthcare contexts or populations. In addition, these models may also overlook more subtle nuances in language that could be important for understanding complex veteran health system research topics. Moreover, the translation of these quantitative findings into actionable healthcare strategies requires careful consideration of context and the diverse needs of the veteran population, which may not be fully captured through this methodological approach. Future research should aim to integrate a wider array of data sources and incorporate qualitative assessments to enrich the analysis and mitigate these limitations.

Conclusion

This project demonstrates the potential of NLP techniques in analyzing VA Health Systems Research Forum publications to uncover meaningful trends and sentiments in veteran healthcare discussions. By implementing and comparing baseline, LSTM, and BERT models, the study revealed significant insights into the evolution of key topics such as telehealth, PTSD, and caregiver. The BERT model outperformed other models in accuracy and topic coherence, while the LSTM model effectively identified temporal trends. These findings highlight the potential for tailored NLP applications to enhance the understanding and accessibility of health information for veterans, informing policy development and healthcare delivery strategies.

References

Carrell, D. S., Halgrim, S., Tran, D. T., Buist, D. S., Chubak, J., Chapman, W. W., & Savova, G. (2014). Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. American journal of epidemiology, 179(6), 749-758.

Houston, T. K., Volkman, J. E., Feng, H., Nazi, K. M., Shimada, S. L., & Fox, S. (2013). Veteran internet use and engagement with health information online. Military Medicine, 178(4), 394-400.

Mitra, A., Pradhan, R., Melamed, R. D., Chen, K., Hoaglin, D. C., Tucker, K. L., ... & Yu, H. (2023). Associations between natural language processing–enriched social determinants of health and suicide death among US veterans. JAMA network open, 6(3), e233079-e233079.

Oliveira, C. R., Niccolai, P., Ortiz, A. M., Sheth, S. S., Shapiro, E. D., Niccolai, L. M., & Brandt, C. A. (2020). Natural language processing for surveillance of cervical and anal cancer and precancer: algorithm development and split-validation study. JMIR medical informatics, 8(11), e20826.

Swed, O., Sheehan, C. M., & Butler, J. S. (2020). The digital divide and veterans' health: Differences in self-reported health by internet usage. Armed Forces & Society, 46(2), 238-258.

Taylor, S., Miller, B. L., Tallapragada, M., & Vogel, M. (2020). Veterans' transition out of the military and knowledge of mental health disorders. Faculty/Researcher Works.

US Department of Veterans Affairs. (2024a). Veterans Health Administration. [Website]. Retrieved from: https://www.va.gov/health/aboutvha.asp#:~:text=The%20Under%20Secretary%20for%20Health,than%209.1%20million%20enrolled%20Veterans.

US Department of Veterans Affairs. (2024b). VA Health Systems Research. [Website]. Retrieved from: https://www.hsrd.research.va.gov/publications/forum/