Understanding Administrative Burden in Medicare and Medicaid: Evidence from Reddit Post Topic Modeling

Sara Murley

December 2025

1 Summary

This topic modeling analysis of Reddit discussions of Medicare and Medicaid reveals that administrative burden arises from both program complexity and fragmented processes. Medicaid-related burden is dominated by eligibility verification, documentation, and caregiving responsibilities, whereas Medicare-related burden primarily involves navigation, cost management, and plan selection.

2 Research Objective

This project aims to analyze public discourse on Reddit surrounding U.S. public health insurance programs, specifically Medicare and Medicaid. In particular, the analysis focuses on how program complexity affects users' understanding, access, and satisfaction.

Medicare and Medicaid are central pillars of the U.S. health care system, but can be extremely difficult to navigate and understand. The burden is placed on beneficiaries to understand multiple plan options, eligibility criteria, coverage limitations, and administrative processes that often vary by state or insurer.

Prior research has highlighted these challenges. Surveys such as the KFF Survey of Consumer Experiences with Health Insurance Lopes et al. (2022), have shown that adults overwhelmingly support public policies to make insurance simpler to understand. Studies of Medicaid administration have also emphasized the importance of consumer engagement in policy design Zhu et al. (2021). Recent work has begun to explore social media as a complementary source of insight for user experiences in public health insurance Chakravarty and Arifuzzaman (2024).

This exploratory study uses text data to identify recurring questions and concerns expressed by users. Insights can help policymakers target areas of confusion and design communication or outreach strategies to reduce barriers to care.

3 Data

This project uses text data collected from Reddit, focusing on posts that discuss Medicare and Medicaid. Reddit organizes conversations into topic-specific communities ("subreddits"). For this analysis, data were collected from **r/Medicare** and **r/Medicaid**.

Reddit data were collected using PRAW, the Python Reddit API Wrapper (Boe, 2025), from publicly available Reddit discussions (Reddit, 2025). The Reddit API documentation can be found here and the PRAW documentation can be found here.

Posts include the full user text along with engagement metadata, such as post score and number of comments. Exploratory data analysis revealed highly right-skewed distributions for both text length and engagement measures (Medicaid distributions shown in Figure 1, with similar Medicare results omitted). To reduce distortion from highly atypical posts while preserving substantive variation, outliers were removed using an interquartile range (IQR) method applied across numeric variables.

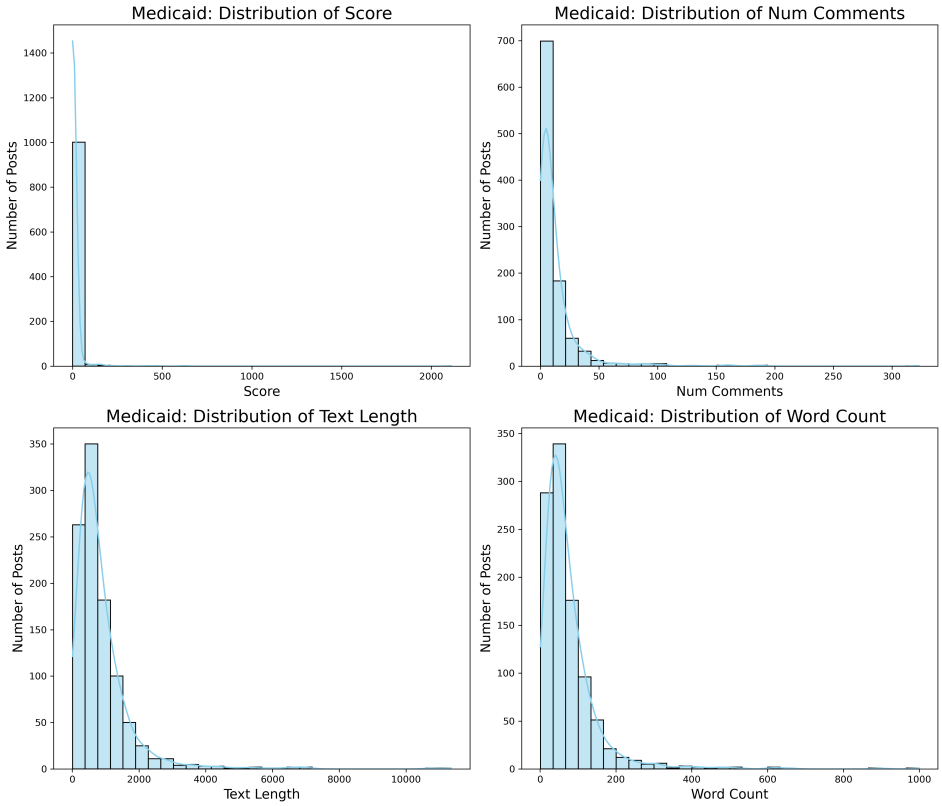

Figure 1: Medicaid Numeric Variable Distributions

After removal of outliers, the final analytic samples included 820 Medicaid posts and 816 Medicare posts. These sample sizes are sufficient for unsupervised text analysis, capturing meaningful thematic variation while remaining computationally tractable.

Reddit functions as an informal peer-support forum where beneficiaries, caregivers, and family members seek advice, share experiences, and troubleshoot administrative challenges. The data consists of user-generated, unstructured text reflecting real-time interactions with public benefit programs rather than institutional narratives or structured surveys.

A preliminary examination of term frequencies confirms the substantive relevance of the corpus. Common phrases such as “need help” and “don’t know” suggest that many posts reflect uncertainty, confusion, or information gaps.

4 Techniques

To identify recurring patterns of administrative burden in Medicare and Medicaid discussions, this analysis applies a structured text-mining pipeline combining natural language preprocessing, TF-IDF feature extraction, unsupervised clustering, and topic modeling. These methods are well suited for exploratory analysis of large, unstructured text datasets where themes emerge organically rather than through predefined labels.

This project builds on prior research that has applied similar pipelines to online health discussions. Ayadi et al. (2023) used NLP and clustering to identify Long COVID symptom categories based on Reddit posts. John and Keikhosrokiani (2022) combined clustering and LDA to classify and understand COVID-19 news content.

Because Medicare and Medicaid differ substantially in program rules, eligibility pathways, and administrative processes, all analyses were conducted separately for each program. This decision avoids conflating structurally distinct administrative experiences and ensures that identified themes are interpretable within each policy context.

4.1 Text Preprocessing and Feature Construction

Before applying clustering or topic modeling, all posts were cleaned and standardized to reduce noise common in social media data, including inconsistent capitalization, informal language, and extraneous symbols. Preprocessing steps included lowercasing text, removing URLs and numbers, filtering stop words, and tokenizing posts into individual words. Contractions were intentionally preserved to reduce distortion of user intent.

Preprocessing choices were validated through iterative inspection of sample posts at each stage, ensuring that meaningful content was preserved while non-informative artifacts were removed. In addition, frequency summaries of tokens before and after cleaning were examined to confirm that administrative terms remained prominent.

The cleaned text was transformed into a numerical representation using term frequency inverse document frequency (TF-IDF). TF-IDF weighting emphasizes words that are distinctive within individual posts while downweighing terms that appear frequently across many posts. This is particularly important for Reddit data, where common words (e.g., "insurance," "plan," "help") can otherwise dominate similarity measures.

To address substantial variation in post length, TF-IDF vectors were L2-normalized. Normalization ensures that similarity comparisons reflect differences in language use rather than differences in verbosity, improving both clustering stability and interpretability. Without normalization, longer posts would exert disproportionate influence on clustering and topic modeling results.

To further prevent generic language from dominating results, document-frequency diagnostics were used to exclude terms appearing in more than 10 percent of posts ($max_df = 0.10$). This threshold was selected based on the observed distribution of token frequencies (example Medicare results in Figure 2) and validated by comparing term lists at alternative cutoffs.

Both unigrams and bigrams were included to capture not only individual terms but also common administrative phrases such as "apply Medicaid," "social security," and "nursing home." The resulting TF-IDF matrices were highly sparse ($> 97\%$), typical for text data. Diagnostic checks confirmed that sparsity did not impede downstream analysis, making additional dimensionality reduction unnecessary at this stage.

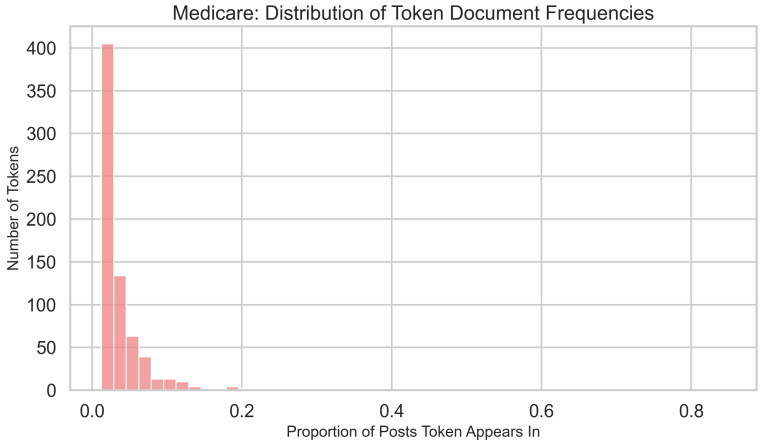

Figure 2: Medicare Token Document Frequencies

4.2 Clustering Techniques

Clustering was used to group posts that share similar language patterns and, by extension, similar types of administrative experiences. K-Means clustering was applied to a combined feature space that included TF-IDF text features and numeric data. All numeric features were standardized prior to clustering to ensure comparability across scales and prevent any single variable from dominating distance calculations. Cluster solutions were evaluated using multiple diagnostics, including elbow plots (within-cluster sum of squares), silhouette scores, and qualitative inspection of top terms and representative posts within each cluster. Figures 3 and 4 show some of these inputs.

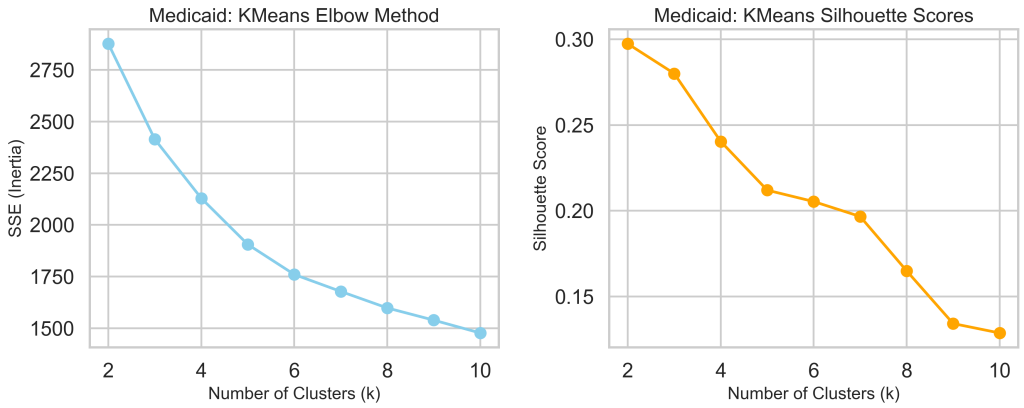

Figure 3: Medicaid KMeans Evaluation and Selection of k

While solutions with fewer clusters maximized statistical separation, they produced overly coarse groupings that masked meaningful variation in administrative experiences.

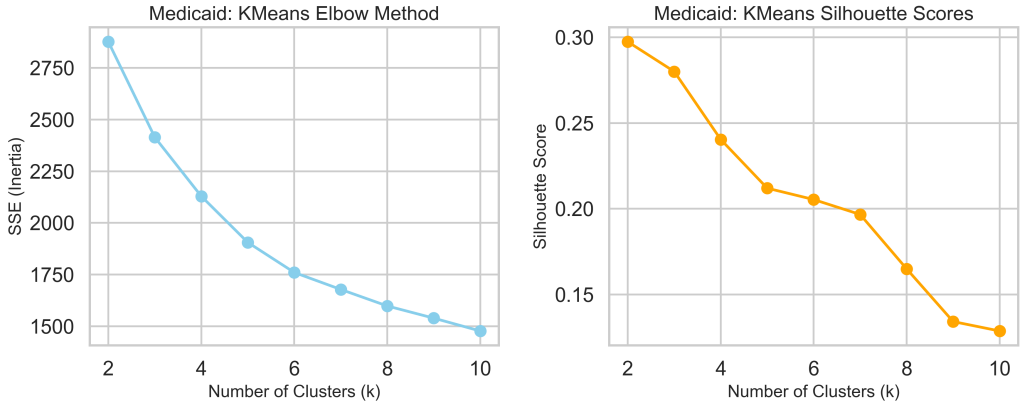

Figure 4: Medicare KMeans Evaluation and Selection of k

A four-cluster solution for Medicaid and a three-cluster solution for Medicare were ultimately selected, reflecting a balance between quantitative fit and substantive interpretability. These choices were validated by examining whether clusters corresponded to coherent administrative themes rather than arbitrary divisions.

4.3 Topic Modeling Techniques

Latent Dirichlet Allocation (LDA) topic modeling was applied using the same vocabulary constraints as the TF-IDF representation. Unlike K-Means clustering, which assigns each post to a single cluster, LDA models each post as a mixture of latent topics, allowing themes to overlap.

Because Reddit posts often reflect multifaceted experiences rather than discrete categories, topic modeling performed better at capturing the underlying structure of the data. Validation focused on interpretability of topic-word distributions, coherence of example posts with high topic probabilities, and stability across alternative topic counts. Correspondence between K-Means clusters and LDA topics was assessed using cross-tabulations.

Together, these techniques provide both structural and thematic insight into how Reddit users experience and articulate administrative burden in Medicare and Medicaid. By validating results at each stage of the pipeline and triangulating findings across methods, the analysis strengthens confidence in the robustness and interpretability of the final insights.

5 Key Findings

5.1 Core Patterns of Administrative Burden

Reddit discussions of Medicare and Medicaid consistently highlight administrative complexity as a major source of burden. While K-Means clustering was initially used to organize posts and guide selection of the number of topics, Latent Dirichlet Allocation (LDA) topic modeling provides the most interpretable and substantive insights, as it allows individual posts to load onto multiple themes. This is particularly important for administrative bur-

den, since users often discuss overlapping issues such as eligibility, caregiving, and costs within a single post. As such, the findings discussed below will primarily focus on the topic modeling results, though readers can find detailed K-Means results in Appendix A.

5.2 Medicaid: Eligibility, Life Circumstances, and High-Stakes Bureaucracy

Topic modeling reveals four dominant themes in Medicaid discussions. The posts are fairly evenly distributed across the 4 topics as shown in Table 1. The top words across topics will be referenced here and can be seen in Figure 5. Distributions showing the engagement and numeric distributions across topics can be found in Appendix A.

Table 1: Number of Posts per LDA Topic for Medicaid

Topic	Number of Posts
0	202
1	225
2	229
3	161

1. Eligibility and application processing

   This topic focuses on general interactions with Medicaid offices and Social Security, particularly around eligibility verification and enrollment. Top words like "cal" and "medi cal" indicate discussions about the California Medicaid program specifically, while "social security" and "ssi" point to coordination with federal benefits programs. Terms such as "office", "called", and "letter" suggest frequent communication with administrative offices, including phone calls and written correspondence.

2. Family-based and caregiving situations

   This topic emphasizes Medicaid issues in family and caregiving contexts. Words like "son", "child", "parents", and "pregnant" show that posts often concern children or dependents. The presence of "denied", "worried", and "applied" highlights frequent concerns about coverage denials, application status, and emotional stress.

3. Accessing Care and Services

   This topic reflects challenges in actually using Medicaid benefits to obtain care. Words like "doctor" and "providers" indicate discussion of accessing medical services. "Hours" and "week" suggest long wait times or limited availability.

4. Long-Term Care and Financial Risk

   This topic concerns nursing home, long-term care, and related financial planning. Words like "nursing home", "long term", and "mother" point to posts about managing care for elderly family members. "Monthly", "bills", and "assets" indicate the financial calculations and documentation required for Medicaid long-term care coverage.

Taken together, these results indicate that Medicaid-related burden is primarily driven by eligibility verification, caregiving and other household circumstances, and financial risk, with repeated interactions required to maintain coverage or access care.

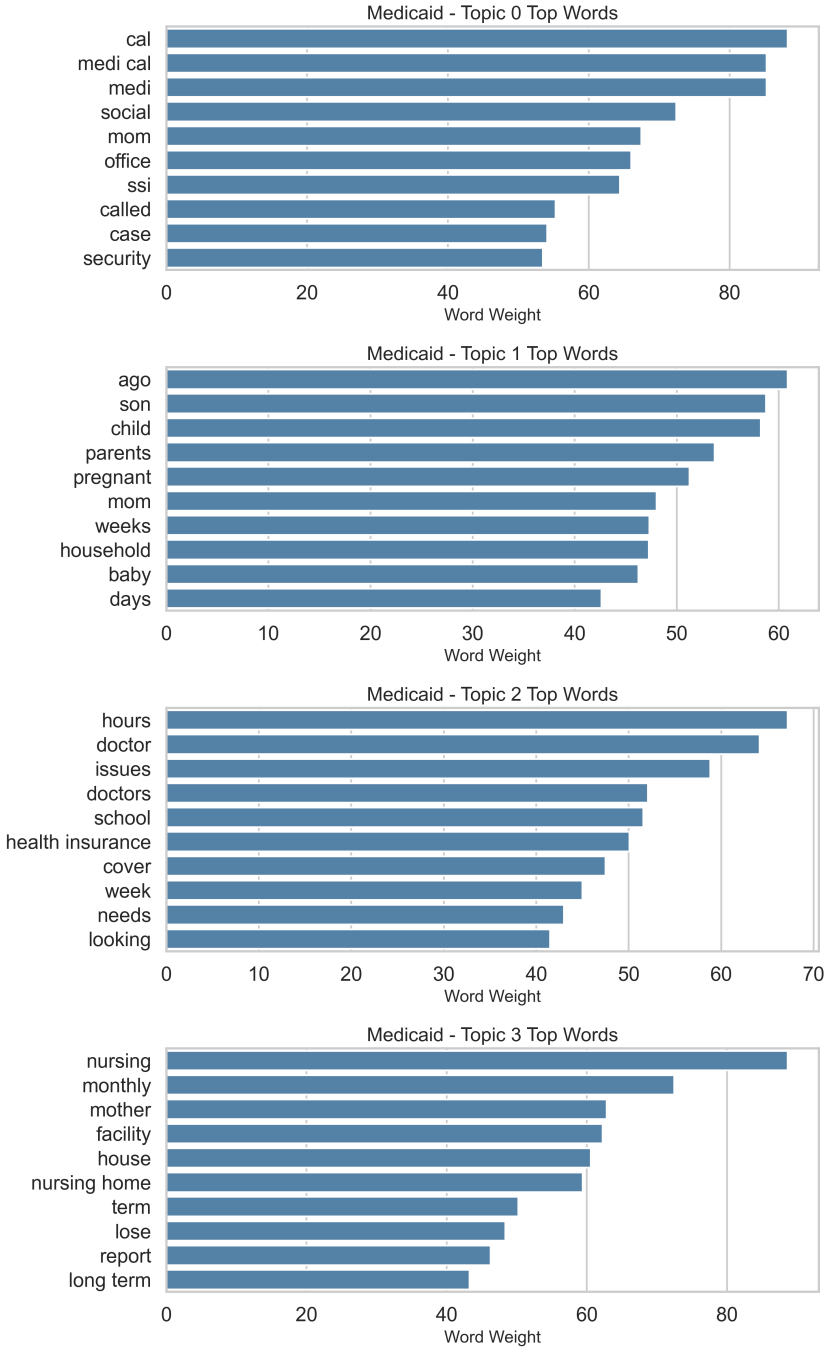

Figure 5: Medicaid Top Words by Topic

5.3 Medicare: Navigation, Cost Management, and Program Fragmentation

Medicare discussions reflect a different form of administrative burden, centered less on eligibility enforcement and more on navigating a complex, fragmented insurance system. Clustering identifies three broad categories of discussion, which are further clarified through topic modeling. Again, the distribution is fairly even across posts (Table 2). Figure 6 shows the top words for each topic, which are summarized below. Details on the numeric variable distributions for each topic can be found in Appendix B.

Table 2: Number of Posts per LDA Topic for Medicare

Topic	Number of Posts
0	297
1	282
2	237

1. Cost management and supplemental coverage

   This topic is centered on managing healthcare costs and navigating supplemental insurance. Words like "deductible", "supplement", and "drugs" indicate discussions about out-of-pocket costs, prescription coverage, and hospital bills. "High" and "paid" reflect concerns about affordability and financial burden.

2. Eligibility, income, and caregiving coordination

   This topic highlights navigating enrollment rules, eligibility requirements, and Social Security interactions. Words like "income", "apply", "qualify", and "enrollment" show that posts often concern determining eligibility or completing the application process. "Social security", "employer", and "work" reflect coordination with federal benefits or employment-based coverage.

3. Plan choice and enrollment decisions

   This topic concerns interactions with Original Medicare, plan networks, and programs that reduce costs for low-income beneficiaries. "Network", "doctor", and "ppo" show that access to providers is a key concern.

These patterns indicate that Medicare-related administrative burden is largely navigation-oriented, arising from information overload and fragmented program design rather than eligibility enforcement. Many of the top terms across multiple topics reflect decisions about which plans to enroll in, highlighting a top source of confusion for Medicare beneficiaries.

5.4 Interpreting Clusters Together

Topic modeling highlights systematic differences in administrative burden between Medicaid and Medicare. Medicaid discussions are dominated by eligibility enforcement, documentation, and high-stakes caregiving decisions. Posts are often triggered by life transitions or financial changes. Medicare discussions emphasize navigation through costs, plan selection, and fragmented coverage, reflecting burdens even for already enrolled beneficiaries.

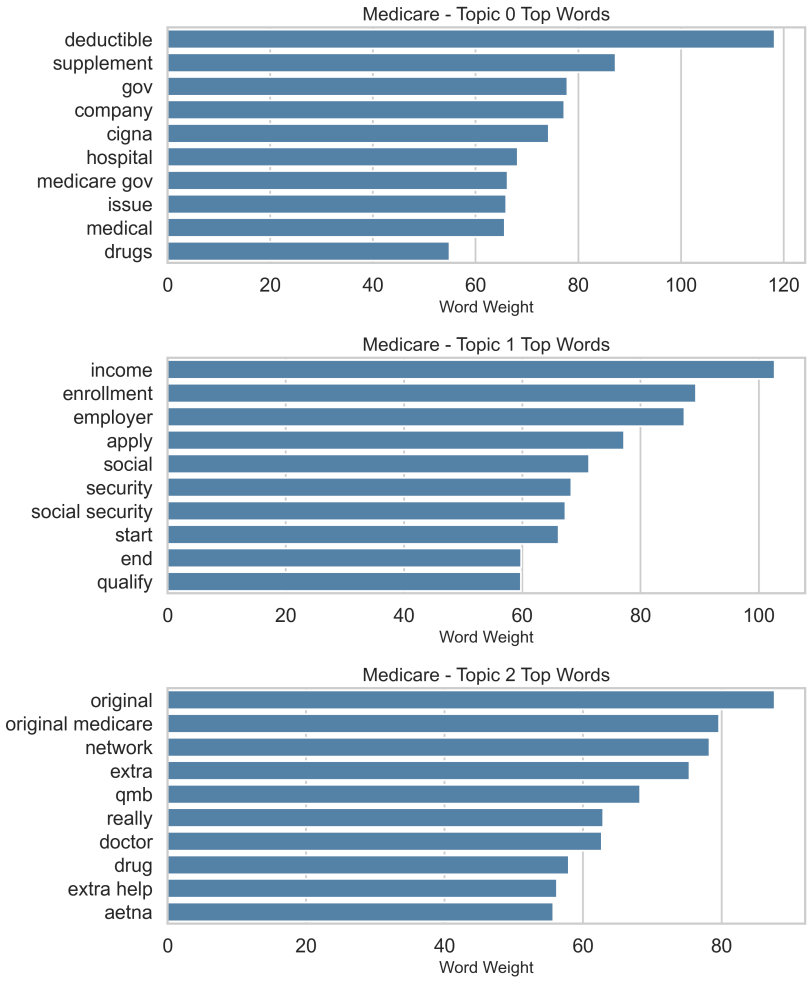

Figure 6: Medicare Top Words by Topic

Despite these differences, both programs impose substantial cognitive, procedural, and emotional effort on users, demonstrating that administrative burden extends beyond clinical care.

K-Means results confirm these patterns: cluster structures correspond broadly to the topics identified in LDA, supporting the robustness of the findings. For readers interested in the detailed clustering results, including top terms and post distributions, see Appendix A.

5.5 Policy Relevant Insights

These findings illustrate how administrative burden manifests in everyday interactions with Medicare and Medicaid. Medicaid discussions emphasize the strain created by eligibility verification, documentation, and coordination across agencies, particularly during periods

of personal instability. Medicare discussions highlight persistent confusion around costs, plan options, and coordination with other programs, even among beneficiaries who are already enrolled.

Rather than pointing to specific policy interventions, these results provide concrete examples of how program design shapes lived experience. Reddit posts reveal where beneficiaries expend time, effort, and emotional energy navigating public insurance systems, offering insight into how administrative complexity can undermine access and satisfaction even when coverage exists.

5.6 Limitations and Future Considerations

This analysis has several important limitations. Reddit users are not representative of the broader Medicare or Medicaid populations; they skew younger, more digitally literate, and more likely to post when encountering problems. As a result, the data likely overrepresent confusion and distress relative to routine interactions.

Posts reflect perceived experiences rather than verified outcomes, and advice shared by other users may be incomplete or inaccurate. Text analysis further abstracts away nuance, and results depend on preprocessing and modeling choices. While multiple techniques were used to assess robustness and interpretability, findings should be understood as descriptive and exploratory rather than causal.

Future work could examine changes in discussion themes over time (e.g., during enrollment periods or across administrations), explore geographic variation in Medicaid experiences, or integrate sentiment analysis to better capture the emotional dimensions of administrative burden alongside procedural complexity.

References

Ayadi, H., Bour, C., Fischer, A., Ghoniem, M., and Fagherazzi, G. (2023). The long covid experience from a patient's perspective: a clustering analysis of 27,216 reddit posts. *Frontiers in Public Health*, 11:1227807.

Boe, B. (2025). Praw: The python reddit api wrapper. `https://praw.readthedocs.io/`. Version 7.7.1 [Computer software].

Chakravarty, U. K. and Arifuzzaman, S. (2024). Sentiment analysis of tweets on social security and medicare. *Social Network Analysis and Mining*, 14(1):91.

John, S. A. and Keikhosrokiani, P. (2022). Covid-19 fake news analytics from social media using topic modeling and clustering. In *Big data analytics for healthcare*, pages 221–232. Elsevier.

Lopes, L., Kirzinger, A., Sparks, G., Stokes, M., and Brodie, M. (2022). Kff survey of consumer experiences with health insurance. *Kaiser Family Foundation. San Francisco.*

Reddit (2025). Reddit posts and comments retrieved via public api. `https://www.reddit.com/dev/api`.

Zhu, J. M., Rowland, R., Gunn, R., Gollust, S., and Grande, D. T. (2021). Engaging consumers in medicaid program design: strategies from the states. *The Milbank Quarterly*, 99(1):99–125.

A K-Means Clustering Figures

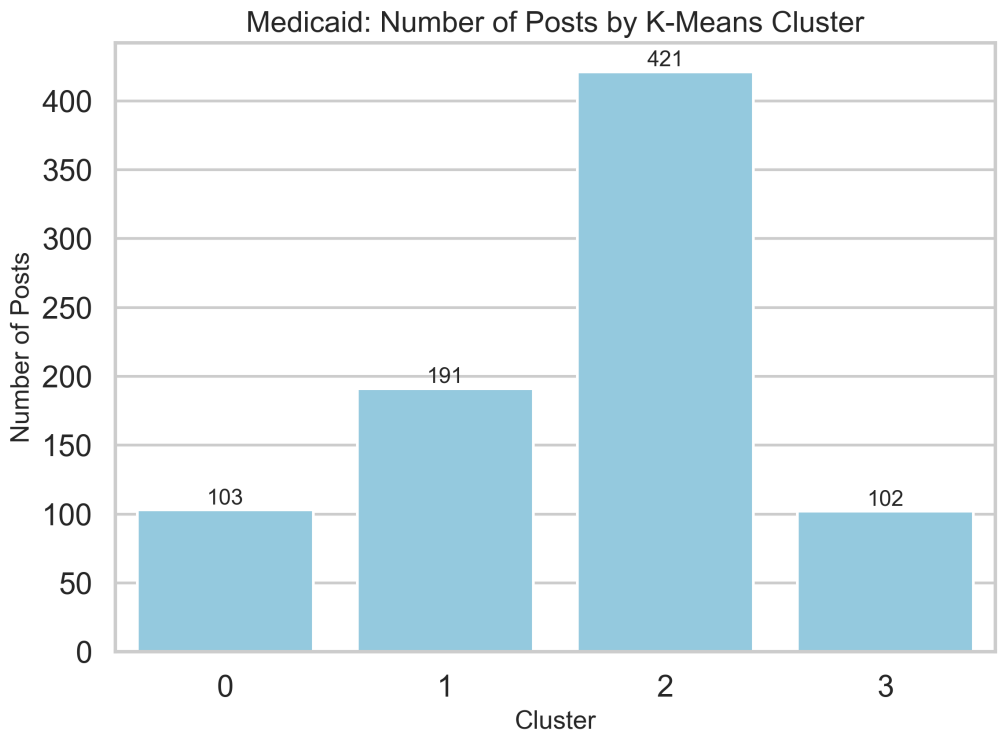

Figure 7: Medicaid Posts by Cluster

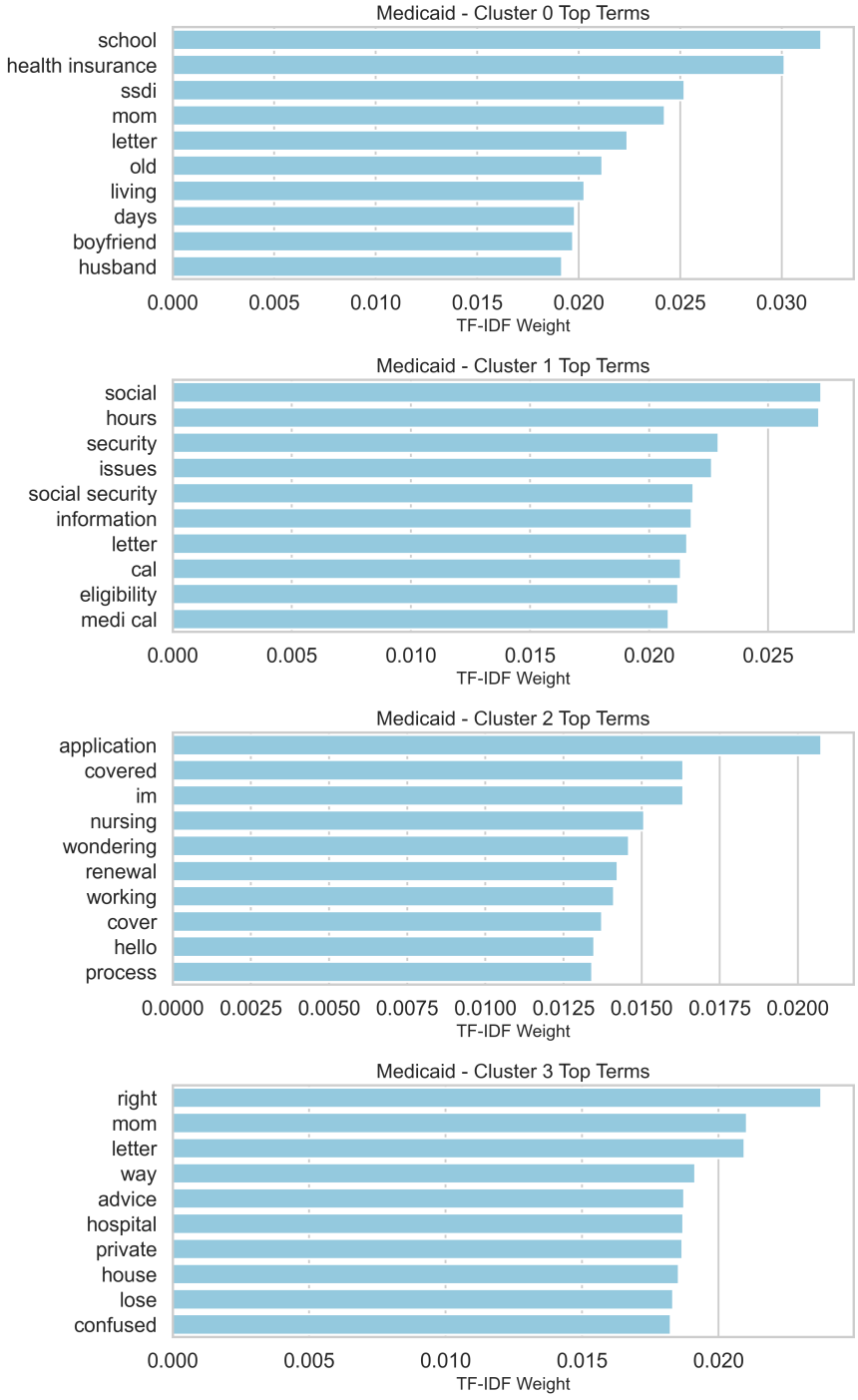

Figure 8: Medicaid Top Words by Cluster

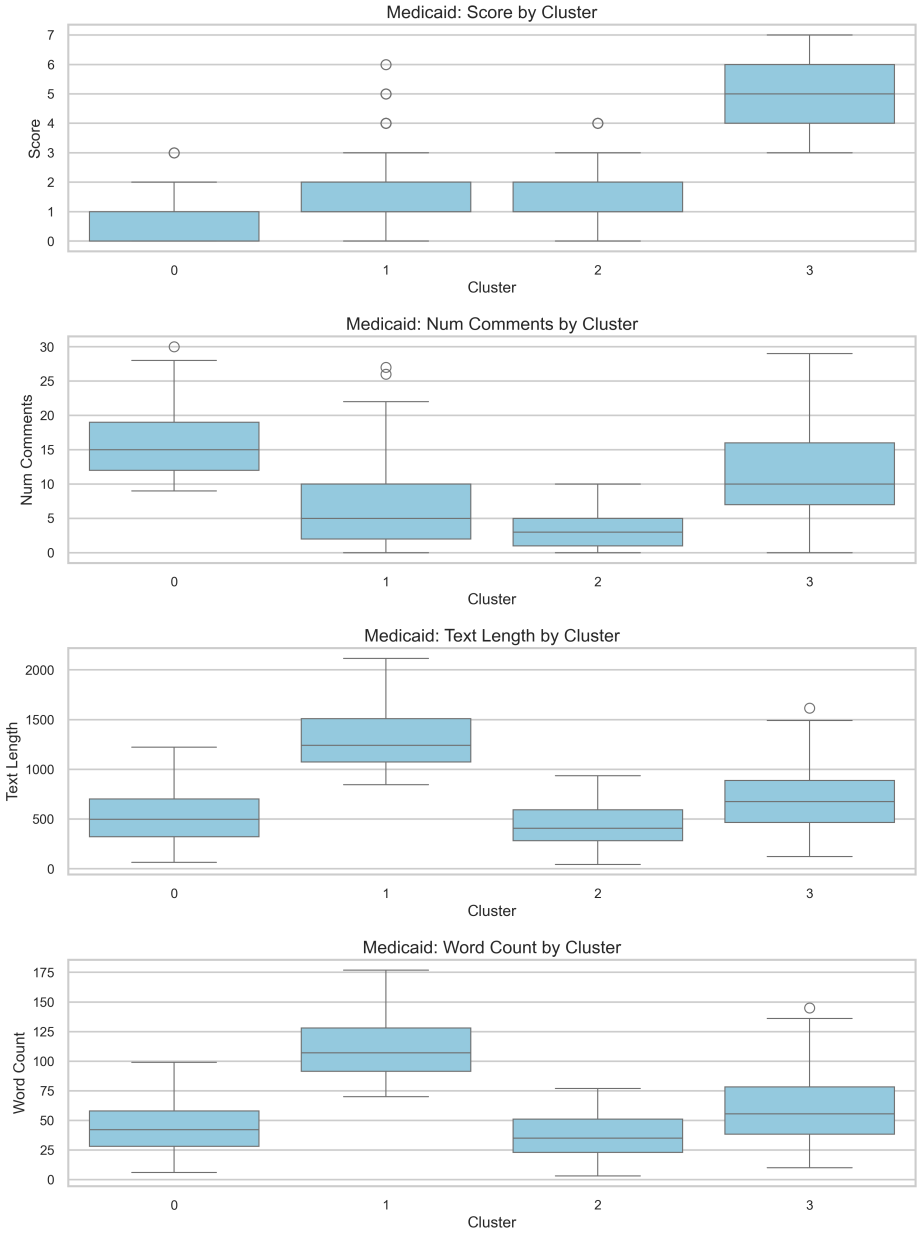

Figure 9: Medicaid Numeric Distributions by Cluster

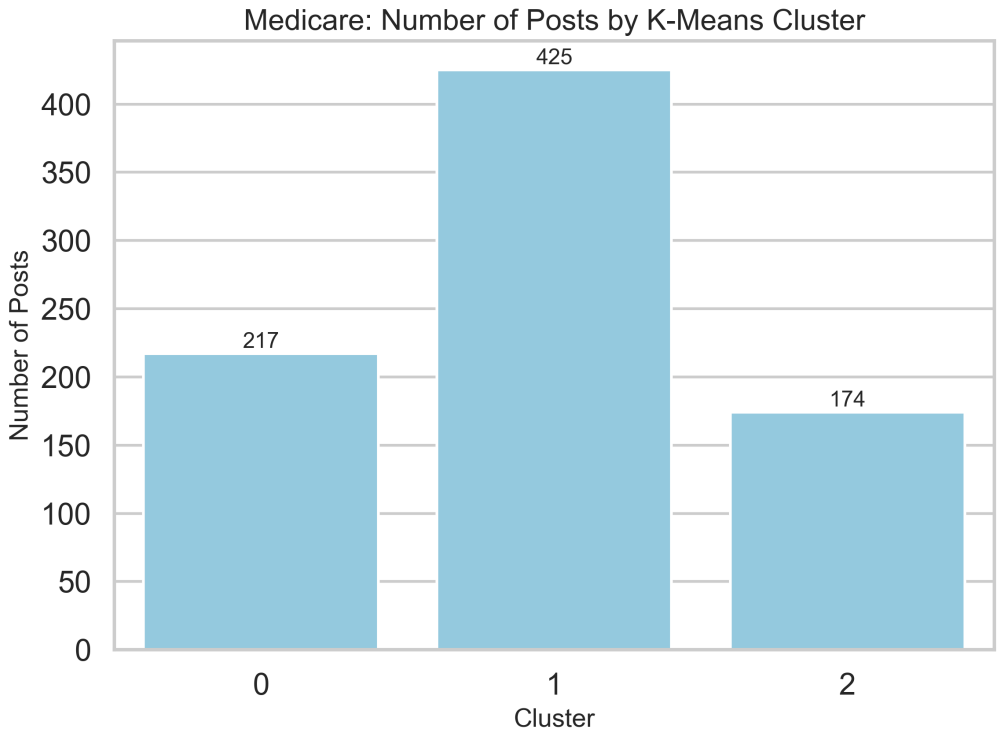

Figure 10: Medicare Posts by Cluster

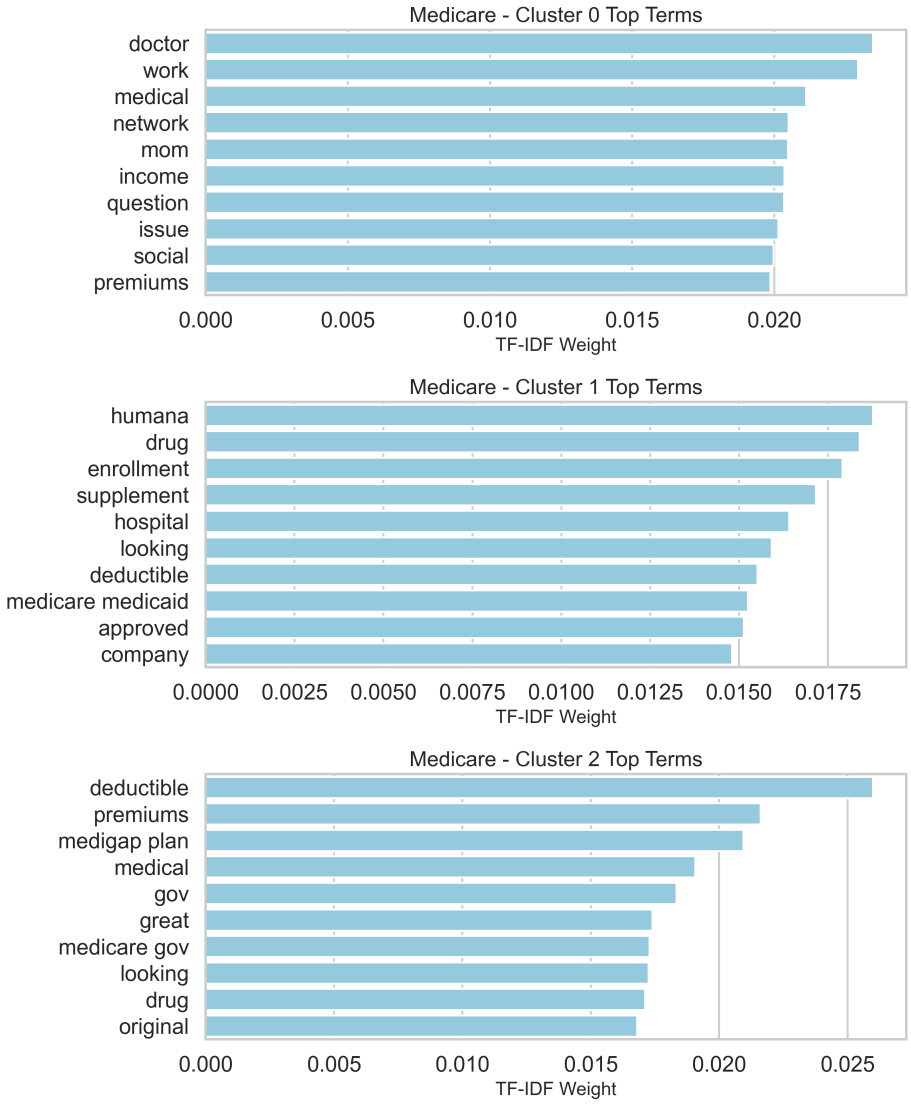

Figure 11: Medicare Top Words by Topic

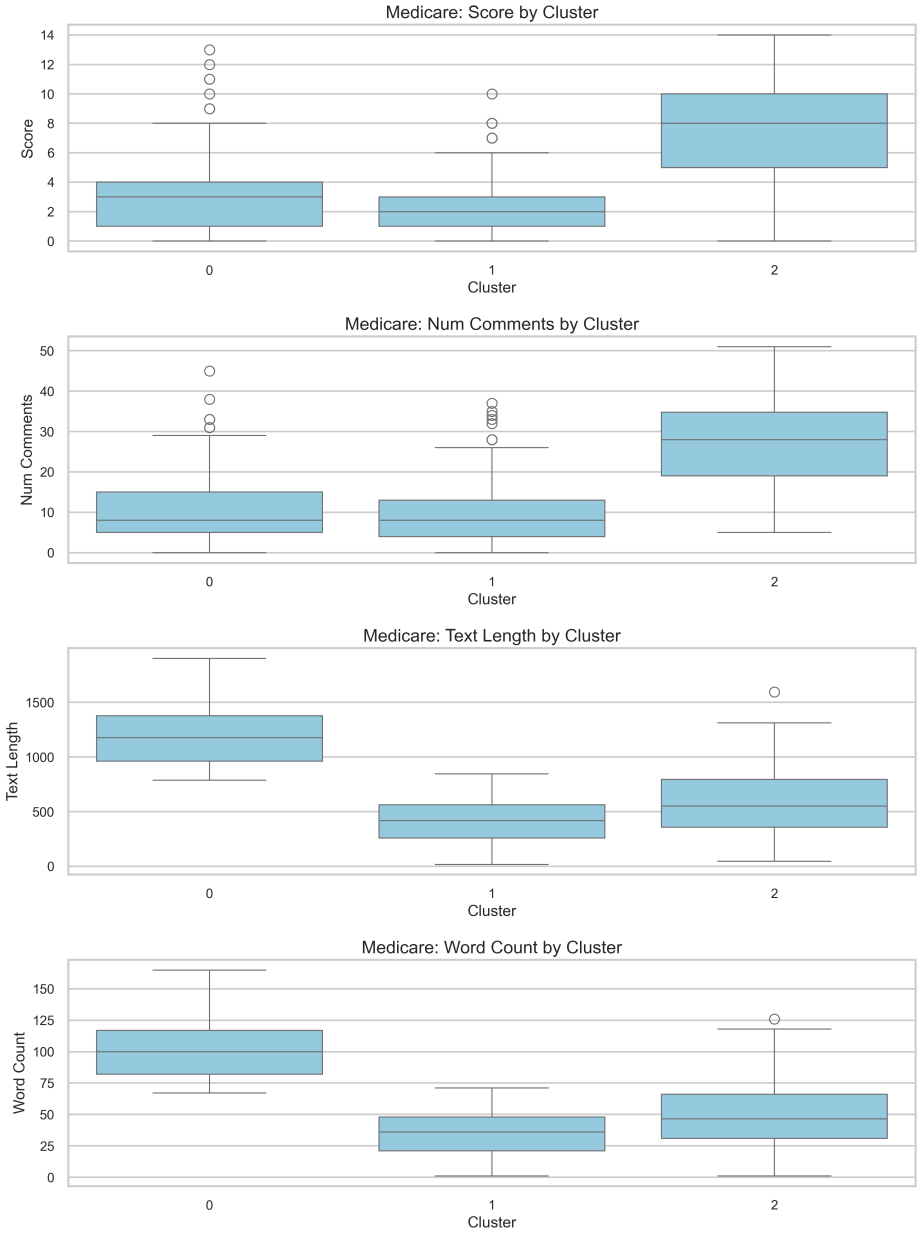

Figure 12: Medicare Numeric Distributions by Cluster

B Engagement and Numeric Distributions by Topics

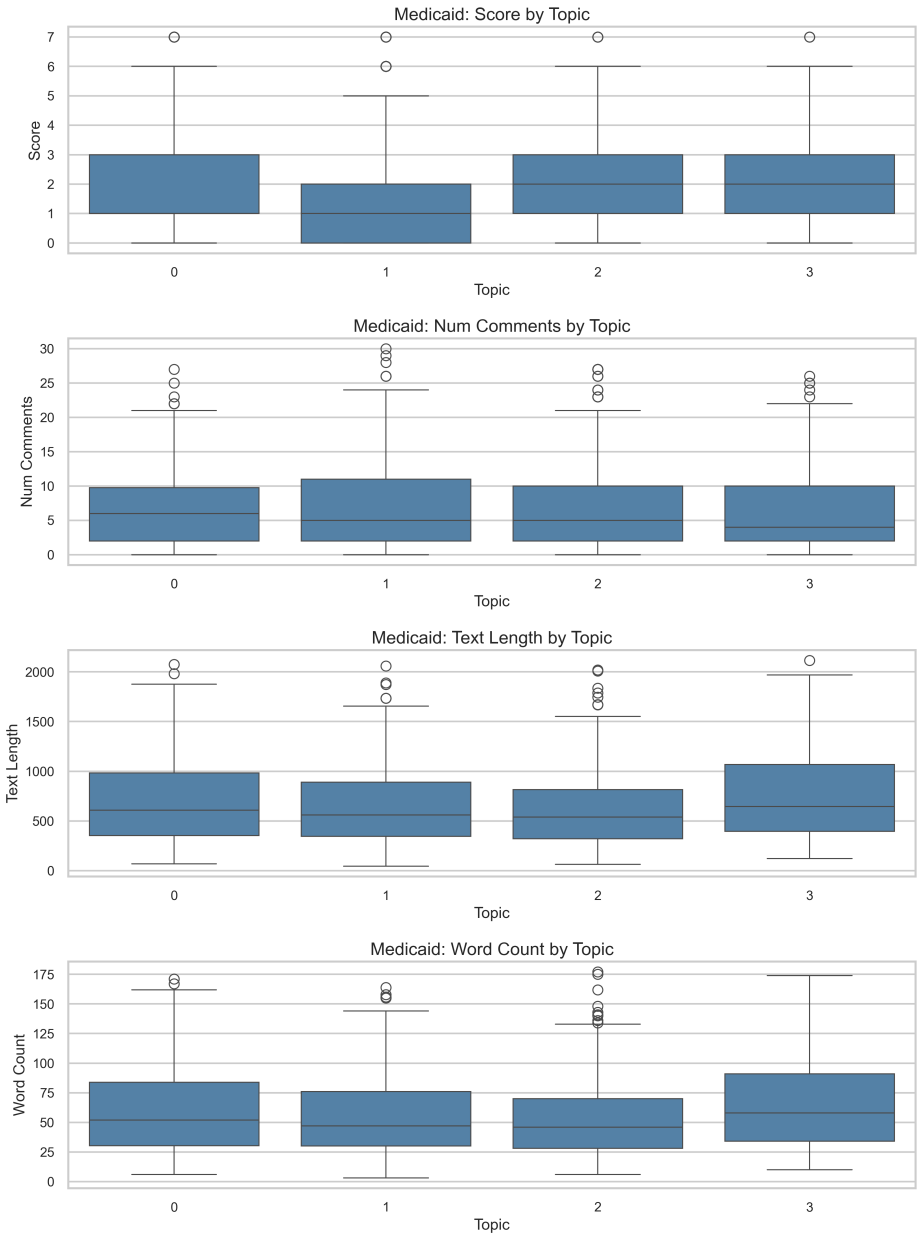

Figure 13: Medicaid Numeric Distributions by Topic

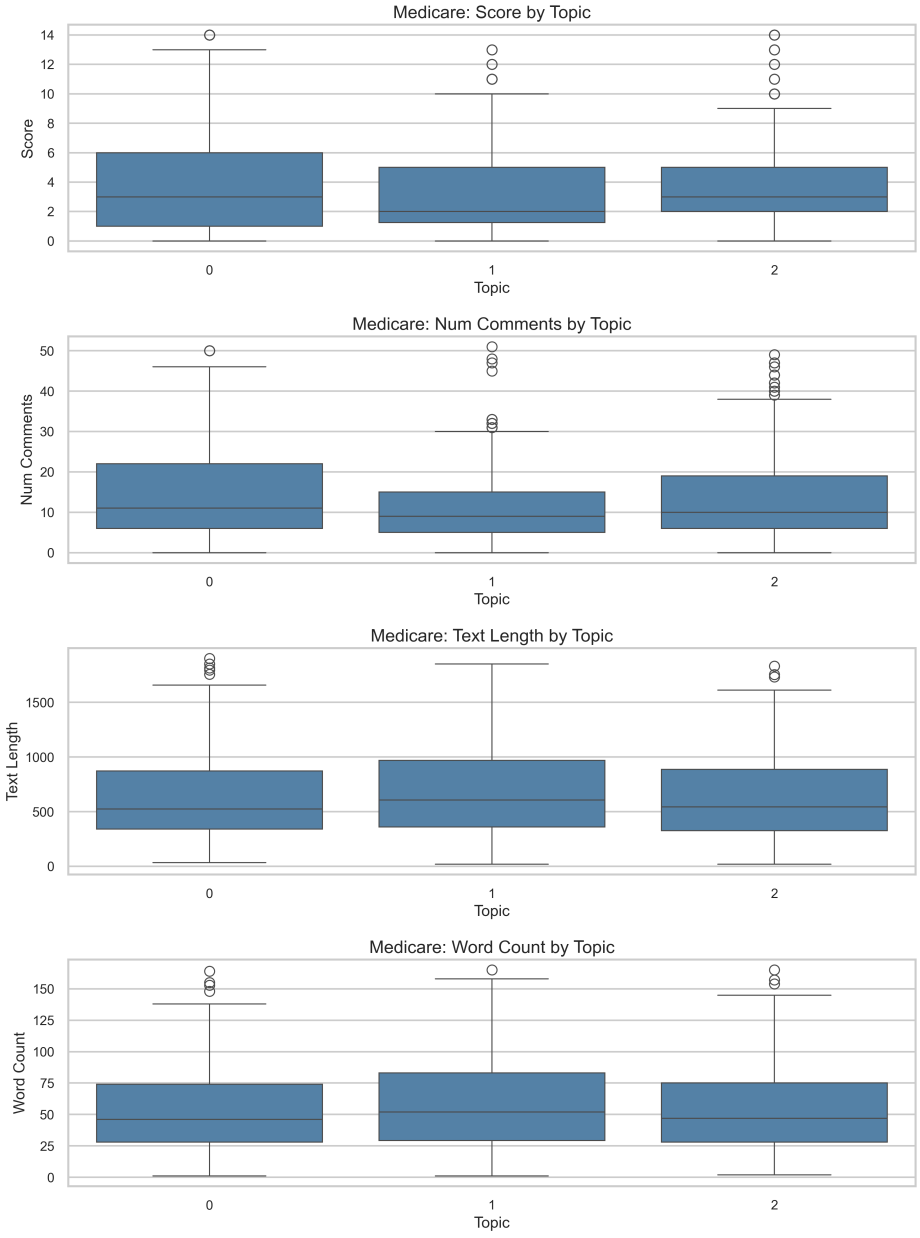

Figure 14: Medicare Numeric Distributions by Topic