Computational Social Sciences (Formative)

Zehao Qian

2024-03-01

1 Topic (Example Analysis)

Topic Name: Examining Mental Health Discourse on Social Media Platforms through Cluster Analysis

The research will explore patterns in mental health-related discussions on various social media platforms, aiming to identify prevalent themes, sentiment, and community engagement around mental health topics.

2 Methods and Dataset

The study will employ cluster analysis (together with QCA and SNA) using a dataset compiled from social media posts related to mental health. These posts can be collected from platforms like X or Reddit (or Some Open Source Dataset), where users frequently discuss personal experiences and opinions related to mental health issues.

A good open source repository on GitHub according to mental health and social media I found is: https://github.com/kharrigian/mental-health-datasets

3 Analysis Plan

3.1 Software

- **Python**: Obtain data from the Internet and develop web crawlers (requests, beautifulsoup and selenium), data preprocess (pandas for data manipulation, scikit-learn)
- **R** for statistics analysis and data visualization (Clustering, SNA, QCA)
- **NLTK** for natural language processing

3.2 Steps

1. Data Collection: Use API tools to gather relevant posts, ensuring to include metadata such as post time, user demographics (if available), and engagement metrics. (Or directly using opensource dataset on GitHub)
2. Data Preprocessing: Clean the data for analysis, including removing irrelevant content, normalizing text, and handling missing values.
3. Feature Extraction: Use natural language processing to extract features from the text, such as term frequency-inverse document frequency (TF-IDF) scores for keywords and phrases.

4. QCA will be applied to categorize social media posts into distinct configurations based on their attributes, such as sentiment, themes, and user demographics. QCA software like fs/QCA can be used for this analysis. The process involves calibrating data into sets, conducting a truth table analysis, and identifying necessary and sufficient conditions for particular outcomes (e.g., high engagement posts).
5. Social Network Analysis will be used to map and measure relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities within the social media discourse on mental health. Using software like Gephi or NetworkX in Python, the analysis will focus on identifying key influencers, information dissemination patterns, and community structures within the mental health discourse network. Metrics such as centrality, density, and clustering coefficients will be calculated to understand the network dynamics.
6. Clustering: Apply clustering algorithms (e.g., K-Means, DBSCAN) to group posts into themes based on similarity in content and sentiment.