A new tool for understanding older people's use of artificial intelligence-powered smart voice assistants: From manual coding to natural language processing coding.

Abstract

Given that cognitive capacities often decline and speech patterns change with age, it can be challenging to understand older people's voice interactions with smart voice assistants (SVA). We compare manual coding and natural language processing (NLP) coding of such interactions and propose a new coding method. English-speaking people 50+ who are living alone have been recruited to interact with Amazon Alexa™ via pre-programmed daily routines for at least 30 minutes daily for three months (*N* = 35; average age = 77.54; average Montreal Cognitive Assessment score = 23.85). Participants' user commands and Alexa™ responses were downloaded as time-stamped, text-based data. A 20% subset (1,020 cases) was first manually coded by human, based on keywords and commands. A rule-based technique was utilized to perform NLP coding of the same subset. A human coder and NLP programmer discussed and resolved discrepancies. The rule-based NLP technique was modified to code the entire data set accordingly. NLP coding is significantly more efficient and less prone to unintentional errors than manual coding. Human feedback can improve NLP coding by providing insights unique to older people, such as speech repetitions or ambiguity, particularly among individuals with diminished cognitive functions.

Keywords: artificial intelligence; smart voice assistants; aging; gerontology; older people; algorithms; natural language processing; coding; content analysis.

Introduction

Older people's adoption of artificial intelligence (AI)-powered smart voice assistants (SVA) has been trending up in recent years favoring speech-based hands-free and eyes-free interactions over typing or clicking modalities [1-3]. To understand older people's SVA use and effects, some researchers rely on qualitative user commands recorded on SVA to uncover detailed insights such as anthropomorphizing SVA or ethical considerations of such AI technology [4-6]. SVA utilizes natural language processing (NLP) AI techniques to interact with user commands. Given that cognitive capacities often decline and speech patterns change with age [7], it can be challenging to understand older people's voice interactions with SVA. We present a comparison of manual coding and NLP coding of such interactions among older people and propose a new method to process and understand older people's voice interaction data with SVA based on data volume, processing speed, accuracy, accountability of context meanings.

Adoption of SVA by Older People

The adoption and use of SVA, such as Amazon Alexa™, Google Assistant™, or Apple Siri™, have gained widespread popularity in recent years. Older people’s adoption of technology in their healthcare management is often aided by factors such as cognitive ability, perceived ease of use, technological self-efficacy and social influence [1-3]. While the hands-free and eyes-free nature of voice interaction with SVA brings the advantages of immediacy and intuition and overcomes physical challenges among older people [1,3], it presents new challenges of understanding older people’s speech interactions with SVA.

Age-related Changes in Cognitive Capacities and Speech Patterns

Older people's cognitive decline is often mirrored by a decrease in the complexity and diversity of linguistic output [8]. Changes in their speech patterns, such as verbal repetitions, pauses, and slowed speech, have been proposed as potential early markers for cognitive impairment [9]. Older people with cognitive impairment, on the other hand, tend to struggle with understanding speech [10]. As such, older people's speech interactions with SVA could be impacted declining cognitive capacities and characterized by repetitions and speech interruptions or irregularities, which can be problematic when processing such qualitative human-SVA speech data.

Coding of Human-SVA Speech Interaction Data

Human-SVA speech interaction data are often manually coded by thematic analysis [11] to identify patterns or themes [3,5]. Yet manual coding of speech data can be prone to potential challenges such as subjectivity, time and labor intensiveness, and human fatigue [12,13], which could be exacerbated by large volumes of textual data generated by human-SVA interaction data. NLP with either human-developed rules or machine learning has been developed to automate the manual coding process by extracting textual phrases evidential of thematic concepts of interest [14,15]. NLP coding of textual data produces rapid results with significant labor- and cost-saving [16,17]. While both NLP methods can automate manual coding, rule-based NLP is better suited for processing human-SVA speech interaction data than machine learning-based NLP based on considerations of sample size. Past studies of older people's use of SVA often have small sample sizes (e.g., less than 100) and usually cannot meet the threshold of adequate effect size for machine learning-based NLP, i.e., 0.5 or higher according to Cohen's scale [18]. In contrast, rule-based NLP can process qualitative textual data from small sample sizes and only needs an NLP expert to develop rules [15]. Accordingly, rule-based NLP is adopted to compare with manual coding by human. The following hypotheses can be proposed.

H1: Rule-based NLP coding of human-SVA speech interaction data will be more efficient than manual coding of the same data.

H2: Rule-based NLP coding of human-SVA speech interaction data will be less prone to unintentional errors than manual coding of the same data.

Materials and Methods

English-speaking people 50+ who are living alone in the United States have been recruited to interact with Amazon Alexa™ SVA via pre-programmed daily routines for at least 30 minutes daily for three months (*N* = 35; 37% male, 63% female; average age = 77.54, range: 50-98). Participants completed Montreal Cognitive Assessment (MoCA) [19] at the beginning of the study (average MoCA score = 23.85, range: 14-30).

Manual Coding

Participants' user commands and Amazon Alexa™ responses were recorded and downloaded from their individual Amazon SVA accounts. The original raw data are time-stamped, text-based commands and responses stored chronologically in 35 individual Excel files. A 20% subset (*n* = 7) of the data (1,020 cases) was first manually coded by human, based on keywords and commands, to sort by weeks and to categorize into 13 different daily routines and other interactions such as entertainment (e.g., music, joke, game/*Akinator*, riddle, meditation/*Five Minute Morning*), information (e.g., weather/*Big Sky*), greetings, calls, settings, and extra (other than pre-programmed daily routines).

Rule-based NLP Coding

Following the procedure outlined in the rule development approach of NLP [15], an expert NLP programmer reviewed the coding standards and developed an understanding of how the keywords and commands were coded. Subsequently knowledge-based rules were utilized to perform NLP coding of the same raw data.

Sorting the data by weeks

To sort the raw data by weeks, various libraries such as pandas, openpyxl, xlsxwriter and Python programming language were used to read and process the same 20% subset (*n* = 7) of the data. The resulting weekly data was then stored in a new DataFrame for easy visualization and interpretation.

Procedure for this rule-based NLP coding of sorting by weeks is outline below.

1. Installation of the abovementioned libraries using pip.

2. Importing required libraries and modules.

3. Reading the Excel file using the pandas(openpyxl) library.

4. Extracting the number of rows and columns in the DataFrame.

5. Defining functions to process the data and calculate weekly time periods.

6. Extracting the start and end times from the DataFrame.

7. Calculating the week stamps based on the start and end times.

8. Creating column names for the new DataFrame representing each weekly period.

9. Constructing a new DataFrame and populating it with the data from the original DataFrame.

10. Handling any remaining data after processing the weeks.

11. Generate the final DataFrame containing the weekly data.

12. Exporting the weekly data to a new Excel file with proper alignment and column width adjustments.

The analysis of the downloaded data using Python resulted in a new Excel file containing the weekly data. The weekly data was organized into columns representing each weekly period, with rows corresponding to different data points or observations of each participant. By using the downloaded data as input for the program, the data is automatically divided into weekly segments within seconds, eliminating the need for any manual intervention.

To validate the accuracy of the data sorting by weeks performed by the program, a comparison was made with the manual coding results based on a randomly selected participant's data. During this comparison, a discrepancy was identified in one instance, specifically a single data cell of week 4 for this participant. According to the NLP program's categorization, it should be classified under week 3 as indicated by the calendar date. Upon discussion with the human coder, it was determined that this discrepancy was an oversight made by the human coder.

Coding the data by categories of routines and other interactions

Using the same 13 categories of daily routines and other interactions from manually coding, the 20% subset dataset was processed using the Python programming language with an XML export feature, leveraging various libraries such as TextBlob and pandas. The TextBlob library proved invaluable in its ability to perform robust text processing tasks, including word extraction and identification of similarities and differences between user commands.

The goal was to utilize the rule-based NLP technique to extract meaningful insights from a dataset consisting of user commands and responses. To achieve this, a set of predefined keywords was established, and various functions were developed to identify these keywords within user inputs. These functions utilized text processing techniques, such as tokenization and word comparison, to determine the presence of keywords and categorize the commands accordingly. The results of the analysis were presented in the form of a comprehensive table, displaying the frequency of each keyword across different weeks. Additionally, an XML export feature enhanced data compatibility and integration.

Details of the rule-based NLP coding by categories of routines and other interactions are described below.

To start, a comprehensive set of predefined keywords was established, representing distinct types of commands. These keywords were organized systematically within an enumerated class referred to as "Keyword," each assigned a corresponding numerical value. This classification system facilitated the straightforward identification and categorization of keywords.

Two crucial functions were developed to identify relevant keywords within user commands: "find_same_words" and "find_different_words." The "find_same_words" function compared pairs of input strings and extracted any words that were identical in both, while the "find_different_words" function identified words present in one string but absent in the other. These functions enabled the identification of repeated commands, ensuring their accurate classification.

Furthermore, additional functions were implemented to specifically detect certain types of commands, such as Music, Weather, Akinator, Calls, and Setting_vol_speed. By utilizing a combination of keyword detection and contextual analysis, these functions determined the relevance and validity of each command, enabling more a focused analysis.

Once the identification and categorization of keywords were completed, a pandas DataFrame named "word_df" was constructed. This DataFrame was structured with columns representing each keyword and rows corresponding to different weeks of user interactions. The frequency of each keyword within each week was recorded, allowing for the tracking of usage trends over time.

Additionally, an XML export feature was incorporated into the methodology. The "word_df" DataFrame was transformed into an XML format using the built-in XML exporting capabilities of the pandas library. This export process preserved the hierarchical structure of the DataFrame, facilitating further data analysis and integration with other tools or systems.

The methodology also involved data aggregation and the calculation of total keyword frequencies. By expanding the "word_df" DataFrame to include a row representing the total frequency of each keyword across all weeks, a comprehensive overview of the most used commands was obtained. This information served as a basis for in-depth analysis of user preferences and behaviors.

In sum, the rule-based NLP coding of categories of routines and other interactions successfully identified, categorized, and analyzed keywords and commands within the user dataset. Utilizing the rule-based NLP technique and custom functions, we extracted meaningful insights and tracked usage trends. The implementation of predefined keywords and functions such as "find_same_words" and "find_different_words" enabled accurate classification of commands, focusing on specific types. The XML export feature enhanced data compatibility and integration.

Results

Inter-coder Reliability Analysis

To ensure consistency and agreement among the human coder and the rule-based NLP programmer in categorizing the commands, inter-coder reliability analysis was conducted. Typically, Cohen's kappa coefficient is used as a measure of inter-coder reliability. However, in this project, there is no predefined definition of correct and incorrect coding, making Cohen's kappa coefficient unsuitable for our dataset.

Instead, we employed the percentage agreement measure to assess the agreement between coders. The percentage agreement calculates the percentage of agreement between coders by dividing the number of agreements by the total number of coding instances and multiplying by 100. While the percentage agreement provides a straightforward measure of agreement, it does not take into account the agreement expected by chance.

The human coder and the rule-based NLP programmer discussed and resolved discrepancies. Based on knowledge learned from the discussion, the rule-based NLP technique was modified to code the entire data set among all 35 participants accordingly.

Hypothesis Testing

H1: Rule-based NLP coding of human-SVA speech interaction data will be more efficient than manual coding of the same data.

Manual coding took an average of 3 hours to code one participant (ranging from 2-5 hours) while the MRLHF method took 9 hours to process 35 participants. Therefore, H1 was supported.

H2: Rule-based NLP coding of human-SVA speech interaction data will be less prone to unintentional errors than manual coding of the same data.

The level of agreement by manual coding and the MRLHF method is 85.49% (872 cases) among the 20% subset data. Of the 14.51% discrepancies (148 cases), 0.29% were program errors (3 cases) and 7.68% were human errors (67 cases) while the remaining 6.54% (78 cases) were results of different coding definitions, e.g., coding repeated commands for clarification by either Amazon Alexa™ or those participant who were associated with high MoCA scores (average MoCA score = 26.2), i.e., declining cognitive functions. As such, H2 was supported.

Discussion

NLP has emerged as a powerful tool for analyzing and processing human language data. As indicated by the results in this study, when applied to examining older people's speech interactions with SVA, rule-based NLP coding offers benefits over manual coding, including increased efficiency and reduced unintentional errors. In the present study, NLP algorithms processed text data at a much faster pace compared to manual annotation. It's also worth noting that the rule-based NLP algorithms can keep processing theoretically unlimited amount of additional data within seconds after the rules have been finalized. In contrast, human coders will always need additional time to code data from new participants. Therefore, like NLP's scalability in business applications [20], the time- and cost-saving of rule-based NLP coding of older people's speech interactions with SVA is exponential. This will free up time and resources for researchers to focus more on discovering insights and helping older people aging well.

Our data also supported the notion that NLP coding can reduce unintentional errors compared to manual coding. The human coder made the unintentional errors because of fatigue and the repetitive nature of processing unstructured text data downloaded from SVA. Coder fatigue is a common threat to the trustworthiness of content analysis when the coding is repetitive and clerical [21]. Our study has provided a solution to prevent such mistakes in future studies of text-based human-SVA interaction data.

Nevertheless, human feedback has provided valuable recommendations to amend the rules for NLP algorithms in the present study. When the human coder and the rule-based NLP programmer discussed their discrepancies on the data by categories of routines and other interactions, coding repeated commands for clarification was identified as one of the main sources of divergence. This observation led to the discovery of the context of such speech repetitions, i.e., the relationship between declining cognitive functions and repeated commands by participants. Without human input, such an important contextual meaning of older people's speech interactions with SVA could have been lost. Consistent with previous research [10], the speech interactions between older people and SVA might have been influenced by cognitive decline, leading to repetitions in their speech patterns. These characteristics pose challenges when processing qualitative human-SVA speech data.

From a user experience (UX) design perspective, researchers and SVA providers should make changes in their AI algorithms to account for older people's changing speech patterns due to cognitive impairment and subsequently improve the flow of conversation between the user and the SVA. Insights from our study can contribute to the design of more inclusive and accessible technologies for older people.

Limitations

While our investigation of rule-based NPL coding of older people's speech interactions with SVA provides evidence for its efficiency gains and error reduction, it is not without its limitations. Like any other NLP algorithms, our model heavily relies on the quality and representativeness of the training data. Future studies should consider a larger and a more diverse sample. This will allow the NLP algorithms to be more accurate when predicting the aging population that is growing and more diverse than ever. Also, our algorithms are programed to process English-language text data. However, it is possible to modify it to process text data in different languages.

Our data were collected after COVID-19 restrictions have been lifted in most cities in the United States. However, the impact of the COVID-19 pandemic might have persisted during our study. Older people have reported increase usage of and more positive attitude toward digital technology including SVA during the COVID-19 pandemic [21]. It is unknown if such trends will continue in the future. We welcome replications of our study to explore the application of rule-based NLP coding of older people's SVA interactions under different circumstances.

In sum, our proposed rule-based NLP coding method, which we intend to make available to other researchers, is an effective tool for investigating older people's speech interactions with SVA or other AI technology.

References

1. Kim S. Exploring how older adults use a smart speaker-based voice assistant in their first interactions: qualitative study. JMIR Mhealth Uhealth. 2021;9:e20427. doi: 10.2196/20427.
2. Kim S, Choudhury A. Exploring older adults' perception and use of smart speaker-based voice assistants: a longitudinal study. Comput Hum Behav. (2021) 124:106914. doi: 10.1016/j.chb.2021.106914
3. Pradhan A, Lazar A, Findlater L. Use of intelligent voice assistants by older adults with low technology use. ACM Trans Comput Hum Interact. (2020) 27:1–27. doi: 10.1145/3373759
4. Kowalski J, Jaskulska A, Skorupska K, Abramczuk K, Biele C, Kopeć W, et al. Older adults and voice interaction: a pilot study with google home. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. Glasgow (2019). p. 1–6.
5. Jones VK, Hanus M, Yan C, Shade MY, Blaskewicz Boron J, Maschieri BR. Reducing Loneliness Among Aging Adults: The Roles of Personal Voice Assistants and Anthropomorphic Interactions. Front Public Health. 2021;9:750736. doi: 10.3389/fpubh.2021.750736.
6. Zhong R, Ma M. Effects of communication style, anthropomorphic setting and individual difference on older adults using voice assistants in the context of health. BMC Geriatr. 2022 doi: 10.1186/s12877-022-03428-2.
7. Murman DL. The impact of age on cognition. Semin. Hear. 2015;36:111–21. doi: 10.1055/s-0035-1555115.
8. Blazer, D. G., Yaffe, K., & Liverman, C. T. (2015). Characterizing and assessing cognitive aging. Washington (DC): National Academies Press

9. Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., & Pakaski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? importance of changes in language abilities in Alzheimer's disease. Frontiers in Aging Neuroscience, 7, 195. doi:10.3389/fnagi.2015.00195

10. Banovic, S., Zunic, L.J., & Sinanovic, O. (2018). Communication difficulties as a result of dementia. Materia Socio-Medica, 30(3), 221–224. https://doi.org/10.5455/msm.2018.30.221-224

11. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol. (2006) 3:77–101. doi: 10.1191/1478088706qp063oa

12. Adu, P. (2019). A step-by-step guide to qualitative data coding. Oxford: Routledge

13. Saldaña, J. (2016). The coding manual for qualitative researchers. Los Angeles: SAGE.

14. Brent, E., & Slusarz, P. (2003). "Feeling the beat": Intelligent coding advice from metaknowledge in qualitative research. Social Science Computer Review, 21(3), 281-303. http://dx.doi.org/10.1177/0894439303253975

15. Crowston, K., Liu, X., & Allen, E. E. (2010). Machine learning and rule-based automated coding of qualitative data. proceedings of the American Society for Information Science and Technology, 47(1), 1-2.

16. Chen, N., Drouhard, M., Kocielnik, R., Suh, J., & Aragon, C. (2018). Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. ACM Transactions on Interactive Intelligent Systems. 8. 1-20. 10.1145/3185515.

17. Evers, J.C. (2018). Current issues in qualitative data analysis software (QDAS): A user and developer perspective. The Qualitative Report, 23(13), 61–73.

18. Rajput, D., Wang, WJ. & Chen, CC. Evaluation of a decided sample size in machine learning applications. BMC Bioinformatics 24, 48 (2023). https://doi.org/10.1186/s12859-023-05156-9

19. Nasreddine Z. S., Phillips N. A., Bédirian V., Charbonneau S., Whitehead V. (2005). The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. J. Am. Geriatr. Soc. 53 695–699. 10.1111/j.1532-5415.2005.53221.x

20. Olujimi, P.A., Ade-Ibijola, A. NLP techniques for automating responses to customer queries: a systematic review. Discov Artif Intell 3, 20 (2023). https://doi.org/10.1007/s44163-023-00065-5

21. Kleinheksel AJ, Rockich-Winston N, Tawfik H, Wyatt TR. Demystifying Content Analysis. Am J Pharm Educ. 2020 Jan;84(1):7113. doi: 10.5688/ajpe7113. PMID: 32292185; PMCID: PMC7055418.

22. Sixsmith A, Horst BR, Simeonov D, Mihailidis A. Older People's Use of Digital Technology During the COVID-19 Pandemic. Bull Sci Technol Soc. 2022 Jun;42(1-2):19–24. doi: 10.1177/02704676221094731. PMCID: PMC9038938.