Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 12, December : 2023

REVIEW ON NATURAL LANGUAGE PROCESSING TECHNIQUES FOR QUALITATIVE RESEARCH

SMD SHAFIULLA[a&b]
[a)] *Assistant Professor*
[b)] *Research Scholar*
Department of Computer Science and Engineering,
[a)]*Scient Institute of Technology, Hyderabad, [INDIA].*
[b)]*Bharatiya Engineering Science & Technology Innovation University (BESTIU), Gorantla, [INDIA].*

ABSTRACT:
Natural Language Processing (NLP) has emerged as a critical technology in the field of artificial intelligence, enabling computers to understand and process human language. Natural Language Processing Tools are essential for various applications, including sentiment analysis, language translation, chatbots, and information extraction. As the demand for NLP continues to grow, numerous tools and frameworks have been developed to facilitate NLP tasks. In this article i am going to present various NLP techniques used in qualitative research. Qualitative methods analyze contextualized, unstructured data. These methods are time and cost intensive, often resulting in small sample sizes and yielding findings that are complicated to replicate. Integrating natural language processing (NLP) into a qualitative project can increase efficiency through time and cost savings; increase sample sizes; and allow for validation through replication.

Keywords : NLP techniques, qualitative research.

I INTRODUCTION

Qualitative data-analysis methods provide thick, rich descriptions of subjects' thoughts, feelings, and lived experiences but may be time-consuming, labor-intensive, or prone to bias. Natural language processing (NLP) is a machine learning technique from computer science that uses algorithms to analyze textual data. NLP allows processing of large amounts of data almost instantaneously. As researchers become conversant with NLP, it is becoming more frequently employed outside of computer science and shows promise as a tool to analyze qualitative data in public health. This is a proof of concept paper [1][2]to evaluate the potential of NLP to analyze qualitative data. Specifically, we ask if NLP can support conventional qualitative analysis, and if so, what its role is. We compared a qualitative method of open coding with two forms of NLP, Topic Modeling, and Word2Vec to analyze transcripts from interviews conducted in rural Belize querying men about their health needs. All three methods returned a series of terms that captured ideas and concepts in subjects' responses to interview questions. Open coding returned 5–10 words or short phrases for each question. Topic Modeling returned a series of word-probability pairs that quantified how well a word captured the topic of a response. Word2Vec returned a list of words for each interview question ordered by which words were predicted to best capture the meaning of the passage. For most interview questions, all three methods returned conceptually similar results.

NLP may be a useful adjunct to qualitative analysis. NLP may be performed after data have undergone open coding as a check on the accuracy of the codes.

II LITERATURE SURVEY

The global increase of data has resulted from the expanded incorporation of electronic devices and software, with increasingly accelerated use over the last 30 years. Quantitative data are produced from everyday items: cars, televisions, electronic medical devices, smart watches, and a myriad of other devices and processes. In 2020, the data volume in bytes exceeds 40 times the number of stars in the observable universe at 44 zettabytes (1,0007 bytes) and with increased production, new and re-emerging analytic methods are deployed to increase our understanding of the world (Desjardins, 2019). Because of a drive to understand the world through data, new statistical techniques are emerging, and older known methods are being given new use. As the amount of qualitative and unstructured data is increasing, developing new methods or integrating methods from other disciplines could expand the capacity of qualitative researchers to analyze larger amounts and more diverse data types. Linking qualitative and quantitative researchers and methods has the potential to expand our understanding of the emerging world of data.

Machine Learning (ML) is one of these re-emerging methods which was developed in the 1950s (Samuel, 1959). Depending on the definition of ML used, there range anywhere from four general concepts to thousands of ML approaches. ML is now used as an umbrella concept for statistical methods that process large amounts of data using algorithms developed and coded in a software application by which a computer iterates and refines analyses (or a statistical model) over time, thereby "learning" the best approach to produce more relevant results. Data mining is an example of a common ML approach where an algorithm is developed and employed to look for relationships between the data which are previously unknown ML has been successful recently in drug discovery, evidence based medicine, and multiple medical healthcare treatments (Alsawas et al., 2016; Lee et al., 2019).

Natural Language Processing (NLP) is an example of an ML method. Like traditional analytics, NLP is a constellation of methods categorized under a concept and not a specific application. NLP is used for qualitative data in the traditional sense or additionally for qualitative data which is collected in other formats, including open[5] ended or free form feedback on a customer satisfaction survey, medical provider notes in an electronic medical record (EMR), or a transcript of research participant interviews (Koleck et al., 2019). NLP can be a cumulative methodological process for researchers beginning with the development of baseline functions and knowledge, which can then be extended to include additional data, programming, and analytic techniques (Miller & Brown, 2018). Two basic and essential NLP concepts are corpus and dictionary. While specific definitions vary, a corpus is generally defined as the entire body of text which contains the contextual information on words in the text, while a dictionary is a list of words in the language of analysis. NLP methods provide exceptional opportunities to analyze large amounts of unstructured, qualitative data, however, as the corpus and dictionary are constructed

by people, they are not bias free nor completely objective. Additionally, to study a new content area, each would need to be updated (Guetterman et al., 2018; Koleck et al., 2019). The potential for investigator introduced bias holds true for any analysis in which a person defines the approach and interprets the findings. NLP offers a distinct advantage, the same algorithm will be applied to all the data, this changes most subjective bias to a systematic bias. When the data are electronic, it is then far less cumbersome for review and analysis by additional persons. As with any replicable approach, over time, there should be a decreased bias and the data should approach convergence between observed outcomes and expected outcomes or observed ou comes and the "truth."

III NLP TECHNIQUES IN RESEARCH

Natural Language Processing (NLP) stands as a multifaceted domain, demanding the adept application of various methodologies to effectively analyze and comprehend human language. In the following discourse, we delve into an exploration and elucidation of a diverse array of techniques that find commonplace utilization in the realm of NLP technology, unraveling the intricate tapestry that underlies the seamless processing of linguistic data. The techniques are

a) Tokenization
b) Stemming & lemmatization
c) Morphological segmentation
d) Stop words removal
e) Text classification
f) Sentiment analysis
g) Topic modeling
h) Keyword extraction
i) Text summarization
j) Parsing
k) Named Entity Recognition

This study evaluated the ability to automate components of qualitative analyses and then compare the outcomes of these themes to the original manual process. NLP captured the overall thematic descriptions however, as used, did not provide rich descriptions that described the nuances of the participant experience, for example the what made the phenomena layered and sensitive. This basic, easy to implement approach of NLP would work best with methods such as content analysis. This method is however, also useful to the researcher who is trying to define or describe a phenomenon using a large qualitative or unstructured data source that seeks to generate the findings from the experience of the participant. There are many positive implications of integrating NLP and other ML approaches into qualitative research. Cost and time savings are among the primary benefits. For investigators who choose this method, the use of a thematic or project based dictionary allows for specialized terms to [10][11]increase the accuracy of identification of influence. Another benefit to our approach was that the automated, machine learning based transcription represents a significant cost savings over human transcription. Using NLP in this research the primary concepts were identified in less than 2 minutes of analysis. Had we replicated

this study through standard qualitative approaches, listening to the interviews would have taken 14 hours, in addition to the time to transcribe and analyze the data. The original investigator estimates the time consumed for this project was at a minimum 120 hours. Additionally, the software applications used for this project were free, including Python and Google Sheets. Qualitative software, like quantitative software, can be expensive and may limit an investigator's ability to complete research.

IV CONCLUSION

Modified and new methods are important to analyze the increasing amount of data in the world. Unstructured data, which is often derived from free form fields and a type of qualitative data, is often left under analyzed given the complexity of analysis (Hong et al., 2019). Unstructured health and medical data offer many possibilities to increase our understanding of disease and symptomology, especially in the area of emerging infections or conditions which require a clinical diagnosis as these data are contained in in many areas within health records. The ability to efficiently and cost effectively harness these data do not only represent a great opportunity in the area of qualitative research but also has implications for more traditional qualitative analysis from the novel methods employed for big data.

REFERENCES

[1] Abram, M. D. (2018). The role of the registered nurse working in substance use disorder treatment: A hermeneutic study. Issues in Mental Health Nursing, 39(6), 490–498. https://doi.org/10.1080/01612840.2017.1413462

[2]Alsawas, M., Alahdab, F., Asi, N., Li, D. C., Wang, Z., & Murad, M. H. (2016). Natural language processing: Use in EBM and a guide for appraisal. Evidence Based Medicine, 21(4), 136–138. https://doi.org/10.1136/ebmed-2016-110437

[3]Alsuhaibani, M., Bollegala, D., Maehara, T., & Kawarabayashi, K. I. (2018). Jointly learning word embeddings using a corpus and knowledge base. PLoS One, 13(3), e0193094. https://doi.org/10. 1371/journal.pone.0193094

[4]Atlas.ti Qualitative Data Analysis Shop. (2020). https://atlasti.clever bridge.com/74/purlorder

[5]Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichle allocation. Journal of Machine Learning Research, 3(January),993–1022.

[6]Carminati, L. (2018). Generalizability in qualitative research: A tale of two traditions. Qualitative Health Research, 28(13), 2094–2101. https://doi.org/10.1177/1049732318788379

[7]Desjardins, J. (2019). How much data is generated each day? https://www.weforum.org/agenda/2019/04/how-much-data-is-generate each-day-cf4bddf29f/

[8]Ding, J., Tarokh, V., & Yang, Y. (2018). Model selection techniques:An overview. IEEE Signal Processing Magazine, 35(6), 16–34.

[9]Guetterman, T. C., Chang, T., DeJonckheere, M., Basu, T., Scruggs,E., & Vydiswaran, V. G. V. (2018). Augmenting qualitative textanalysis with natural language processing:Methodological study.Journal of Medical Internet Research, 20(6), e231. https://doi.org/10.2196/jmir.9702

[10] Hong, N., Wen, A., Mojarad, M. R., Sohn, S., Liu, H., & Jiang, G. (2018). Standardizing heterogeneous annotation corpora using HL7 FHIR for facilitating their reuse and integration in clinical NLP. AMIA Annual Symposium Proceedings, 2018, 574–583.

[11] Hong, N., Wen, A., Shen, F., Sohn, S., Wang, C., Liu, H., & Jiang, G.(2019). Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. JAMIA Open, 2(4),570–579. https://doi.org/1010.1093/jamiaopen/ooz056

[12]Koleck, T. A., Dreisbach, C., Bourne, P. E., & Bakken, S. (2019).Natural language processing of symptoms documented in free-text narratives of electronic health records: A systematic review. Jounal of the American Medical Informatics Association, 26(4),364–379. https://doi.org/1010.1093/jamia/ocy173

[13]Konovalov, S., Scotch, M., Post, L., & Brandt, C. (2010). Biomedical informatics techniques for processing and analyzing web blogs of military service members. Journal of Medical Internet Research,12(4), e45. https://doi.org/10.2196/jmir.1538 from https://www.spyder-ide.org