2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Designing a Natural Language Processing System to Support Social Science Research

Keshava Pallavi Gone
Interdisciplinary PhD Candidate
Dalhousie University, Canada
keshavapallavi.gone@dal.ca

Supervisor: Michael Smit
Professor, Information Management
Dalhousie University, Canada
mike.smit@dal.ca

Abstract— **The rapid development of machine learning has delivered new approaches, methods, and tools to multiple domains. I see potential for these developments, specifically natural language processing (NLP), to provide new insights, novel methods, and larger scale to social science research. However, novel NLP methods require substantial technical skills to implement. Some of the highest adoption of novel technical tools is in the area of social media analysis, where the volume of source material can overwhelm methods that rely on human capacity. My PhD dissertation aims to bridge the gap between NLP technologies and the unique needs of social science research by contributing to the development of an open-source NLP tool specifically tailored for social science researchers that reduces barriers to entry. The goal is to empower social science researchers by providing more opportunities to explore data in novel ways. This paper outlines the objectives, methodology, and expected outcomes of the proposed research study, which includes designing the development process, requirement analysis, prototyping an NLP tool, evaluating its usability and performance, and providing support for its integration into the research workflow.**

Keywords—Natural Language Processing, Text mining tool, social science research

I. Introduction

NLP techniques can help researchers find insights in massive data sets and can provide a unique lens on data by identifying patterns, keywords, or trends in text corpora. There are several tools and libraries that implement text mining techniques, but the barriers to entry remain high and few open-source tools are intended to support social science research.

Researchers in multiple fields in social science have been utilizing NLP text mining techniques for various applications. Methods like text preprocessing, clustering, classification, topic analysis and sentiment analysis are assisting researchers to examine values, cultures, sentiments, public behaviour, and relationships [1][2][3][4]. Though social scientists are analyzing text using NLP to reveal valuable insights, they still face challenges in implementation.

One of the major challenges is the need for programming skills for using such advanced technologies [5][6][7]. Researchers without technical or data science skills are not able to use an NLP system for implementing data analysis using machine learning models. While people with technical skills can be hired, they lack the domain knowledge required. Often, social science researchers learn enough technical skills to perform the analysis, which is time-consuming and requires mentorship that is in short supply in the field.

A related challenge is these researchers are not equipped to develop general systems: they do enough to complete their specific task, but the code is not re-usable or adaptable for other applications. This is certainly the case in our group's previous experience, which focused on using novel image processing methods to assess social impact (e.g. [8]), and in our review of natural language processing tools in a social science context [9].

A final challenge is that while commercial tools exist to provide automated text analysis, they often focus on supporting existing methods in social science rather than allowing space for developing novel methods and are not widely adopted due to concerns about the cost and the lack of transparency [10].

An open-source NLP tool is essential for social scientists that do not have a programming background to analyze high-quality and high-quantity data. It will make advanced data analysis, text mining, and information extraction available widely to social science researchers, for any who wish to consider these methods. Social science researchers will be able to use the tool to develop novel methods, derive broader insights, improve research efficiency, and make evidence-based contributions to research in the field.

II. Objective of the Study

The overall objective of the study is to bridge the gap between modern machine learning and social science research by contributing to the development of a dedicated open-source NLP tool prototype that overcomes social scientists' big data

ASONAM '23, November 6–9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11...$15.00
https://doi.org/10.**1145/3625007.3627479**

issues. This will require achieving the following research objectives:

i. To identify and examine social scientists' requirements and expectations for an NLP system.

ii. To design and develop the prototype of the NLP tool using a user-centered design approach, including comprehensive support for its integration into social science research workflows.

iii. To evaluate the usability and performance of the tool.

The first objective includes a literature review of how NLP is applied in social science research, which I will use to inform a User-Centered Design (UCD) approach that includes user requirement analysis to examine researcher needs, expectations and opportunities. The second objective builds on the first, continuing an iterative UCD approach to develop a system prototype and provide support for integrating it in research workflows. Finally, the tool's usability and performance will be evaluated and refined based on user feedback.

III. Data Collection And Methodology of Use

A. Requirement Analysis and Gathering

User Centered design approach is used in developing an NLP tool, as it allows the users to involve early in the development process through the designing and testing of the technology. Fig. 1 presents an overview of the tasks involved in each stage of the UCD process for developing the NLP system. The initial step in UCD is the requirement analysis, the process collects the necessary information needed to design the system. To identify the user needs and expectations for the NLP text mining tool, social science researchers will be recruited to gather their needs, limitation expectations, and comfort with technology in the context of machine learning to support social science. User interviews will be analyzed to understand the existing approaches, challenges, and opportunities that comprise user needs.

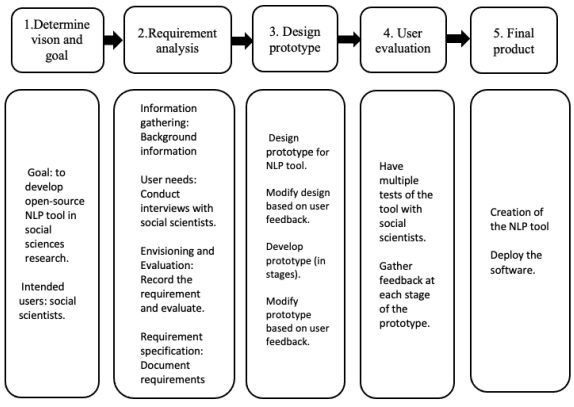

Fig. 1. User centered Design for NLP System

a) Recruitment: My recent research study (Gone et al., 2023) is a comprehensive literature review on the applications of NLP text mining techniques in social sciences using social media data. The study reviewed 118 papers from January 2018 to May 2022 that demonstrated the applications and techniques of text mining in various domains of the social science discipline. I will begin recruitment with the first authors of these articles based on data from Scopus, and an email interview invitation will be sent with a description of the study.

b) Requirement analysis: Semi-structured interviews will be conducted to gather requirements from the researchers. A series of questions are prepared to start the conversation during the interviews (see Appendix). The interviews will be conducted online, and the transcripts will be analyzed to synthesize requirements. The questions and approach were developed based on the initial conversations made with social scientists as expert informants. I interviewed four researchers to obtain background information on the research practices that currently takes place in the discipline and their point of view in developing the new open-source NLP tool. The conversations provided a detailed insights into the existing research processes and tools implemented by the researchers in social sciences. The limitations with the existing methods and expectations for implementing NLP methods were also addressed. They need an NLP tool that is free, user-friendly with better interaction with the system and does not require programming knowledge. The valuable inputs from the experts assisted to develop a research plan to conduct requirement analysis. One outcome of this analysis will be to identify specific fields that will benefit from this tool as early adopters; a target audience of social science users is too broad to serve effectively with a prototype.

B. Development and Evaluation of the NLP tool

The NLP tool will be developed using the Python programming language, which is widely used to implement NLP techniques. Based on the requirement analysis, the tool will support specific social science data as input, in addition to general text input. This includes social media posts, online forums, survey, and other relevant text-based data. Again, based on the requirement analysis, NLP techniques to include will be prioritized, but examples might include preprocessing, part of speech tagging, named entity recognition, sentiment analysis, text classification, and topic modelling.

Once a minimum viable product is complete, it will be deployed to user research workflows for testing the usability and performance. User feedback will be collected, and the prototype improved and further developed iteratively.

IV. Expected Outcome

The requirement analysis and user centered design processes will themselves yield valuable insight into the barriers, gaps, and pain points experienced by social scientists when using natural language processing tools. We expect these to include concerns about the transparency of NLP methods, technical skills required, and acceptance in the research community.

The primary outcome is a prototype tool, which will serve the same role as a microscope for a biologist: a user-friendly tool that requires some skill to use, but which primarily relies on the research expertise of the user to be used effectively and can show things not visible to humans without the help of technology. It will address specific needs of social science researchers who want to streamline data analysis and extract valuable insights from textual data by integrating this tool

into their research workflow. It will be assessed based on its usability and performance, and how well it meets elicited requirements.

V. Conclusion

This project will empower social science researchers by providing them with a dedicated tool that enhances their data analysis capabilities and connecting social science research to the fast-moving field of artificial intelligence. It will yield insights into how users imagine and adopt.

Acknowledgment

With thanks to advice from Kate Sherren, Vlado Keselj, and Colin Conrad.

References

[1] T. T. Aurpa, R. Sadik, and M. S. Ahmed, "Abusive Bangla comments detection on Facebook using transformer-based deep learning models," *Social Network Analysis and Mining*, vol. 12, no. 1, Dec. 2021, doi: https://doi.org/10.1007/s13278-021-00852-x

[2] W. Chen, K. K. Lai, and Y. Cai, "Exploring public mood toward commodity markets: a comparative study of user behavior on Sina Weibo and Twitter," *Internet Research*, vol. 31, no. 3, pp. 1102–1119, Nov. 2020, doi: https://doi.org/10.1108/intr-02-2020-0055.

[3] M. A. Faruque, S. Rahman, P. Chakraborty, T. Choudhury, J.-S. Um, and T. P. Singh, "Ascertaining polarity of public opinions on Bangladesh cricket using machine learning techniques," *Spatial Information Research*, vol. 30, no. 1, Apr. 2021, doi: https://doi.org/10.1007/s41324-021-00403-8.

[4] S. Tartir and I. Abdul-Nabi, "Semantic Sentiment Analysis in Arabic Social Media," *Journal of King Saud University - Computer and Information Sciences*, vol. 29, no. 2, pp. 229–233, Apr. 2017, doi: https://doi.org/10.1016/j.jksuci.2016.11.011.

[5] J. Buckley, M. Brown, S. Thomson, W. Olsen, and J. Carter, "Embedding quantitative skills into the social science curriculum: case studies from Manchester," *International Journal of Social Research Methodology*, vol. 18, no. 5, pp. 495–510, Jul. 2015, doi: https://doi.org/10.1080/13645579.2015.1062624.

[6] T. Donoghue, B. Voytek, and S. E. Ellis, "Teaching Creative and Practical Data Science at Scale," *Journal of Statistics and Data Science Education*, vol. 29, no. sup1, pp. S27–S39, Mar. 2021, doi: https://doi.org/10.1080/10691898.2020.1860725.

[7] M. Zwilling, "Big Data Challenges in Social sciences: an NLP Analysis," *Journal of Computer Information Systems*, vol. 63, no. 3, pp. 537–554, 2023.

[8] K. Sherren *et al.*, "Social media and social impact assessment: evolving methods in a shifting context," *Current Sociology*, (in press).

[9] K. P. Gone, M. Smit, K. Sherren, V. Keselj, and C. Conrad, "Applications of NLP in social science research leveraging social media data: A literature review using text mining," unpublished manuscript, 2023.

[10] M. Strobel, "Aspects of Transparency in Machine Learning," *Adaptive Agents and Multi-Agents Systems*, pp. 2449–2451, May 2019.

Appendix

List of questions prepared to discuss during the interviews to gather requirements to design NLP system.

i. What specific research questions or topics do you aim to investigate using text mining? What are the key objectives of the text mining NLP tool you require?
ii. What features or capabilities would be essential for the NLP tool to meet your research requirements.
iii. Can you describe any existing tools or software you have used for text mining and what limitations you experienced with them?
iv. What type of text do you analyze for your research? How much data do you plan to process?
v. What would be the ideal output format for the NLP tool results? Can you provide some examples of the types of insights or patterns you are hoping to uncover through text mining?
vi. Are there any specific domain-specific requirements that need to be considered when designing an NLP tool for social science research?
vii. How easy user-interface is expected to be? Are there any specific UI related considerations that need to be taken into account when designing an NLP tool?
viii. What are the specific metrics or evaluation criteria that social science researchers use to assess the performance or quality of NLP methods in your research?
ix. Based on what you've seen, what advances in the NLP are the most potential to be impactful in social science research? Can you provide any use case or success stories where NLP tools have been instrumental in advancing social science research?
x. Are there any ethical considerations or challenges unique to using NLP in social science research? How do you currently address or mitigate these concerns?
xi. Are there any concerns related to data privacy and data protection when working with sensitive or personal information? How do you think it should be handled?
xii. Is there a risk in using the new open-source NLP tool and would that impact the social science research?