Leveraging User Sentiment Analysis for Continuous Improvement in Scientific Workflow Management Systems

Project Proposal By – Shamse Tasnim Cynthia

Background and Motivation: Sentiment analysis, a pivotal aspect of natural language processing, has drawn significant attention within the software industry [1]. In recent years, developers have increasingly leveraged sentiment analysis techniques to quantify public opinion and sentiments surrounding their software projects, tools, and technologies. By analyzing online discussion forums, and social media instructions related to programming languages, frameworks, development practices or domain-specific systems, developers gain insights into user satisfaction, concerns and emerging trends [2]. This insight allows developers to make informed decisions, prioritize features and refine their products to align with user expectations [3]. The integration of sentiment analysis into developer toolkits showcases its growing significance in understanding user sentiments within the software development landscape. Several studies have been conducted analyzing the sentiments within different targeted communities serving different purposes [4]–[9]. Nevertheless, sentiment analysis within the scientific workflow management system (SWfMS) communities has remained largely unexplored until now.

Consequently, we have redirected our research focus towards investigating the sentiment dynamics within these SWfMS communities. These systems such as Galaxy [10], Taverna [11], Kepler [12], VizSciFlow [13], SnakeMake [14] etc. play a pivotal role in facilitating the efficient analysis of extensive datasets, which is crucial for addressing complex challenges such as plant phenotyping, water quality prediction, food security, global water security, cancer research, and more. With the exponential growth in data volume and the demand for large-scale experiments, users have been compelled to navigate through increasingly complex SWfMSs [15]–[18]. Sentiment analysis can enable a deeper understanding of users' sentiments and experiences with SWfMS. By analyzing user feedback, reviews, and discussions, developers can identify areas that users find intuitive, challenging, frustrating, or beneficial within the SWfMSs. This understanding is crucial for enhancing the overall user experience and usability of the system. Our ultimate objective is to enrich the research experience of scientists by providing them with a user-friendly, dependable, collaborative, and scalable computational environment achieved through a comprehensive analysis of their sentiments concerning these platforms.

Research Questions:

1. **What are the predominant sentiments expressed by users within the scientific workflow management system community, and how do these sentiments vary across different scientific domains or research areas?**
2. **How do sentiments towards specific features or functionalities of scientific workflow management systems impact user satisfaction and adoption?**

3. **To what extent can sentiment analysis be integrated into the software development and evolution processes of SWfMS to optimize feature prioritization, user-centric design, and continuous improvement?**

Methodology: To investigate the lived experience of users on SWfMSs, we will take a qualitative approach that aligns with Hoda's Socio-Technical Grounded Theory, basic stage for data collection and analysis [19]. We will take our sample dataset from different user forums provided by the SWfMSs and also the discussion forums of the software repositories. We will analyze the data initially and present hypotheses to summarize our findings. Then, we will collect more data to conduct the rest of the experiments. We intend to create a rich subset by filtering the whole dataset using a semi-automated approach that employs sentiment analysis to select posts that will contain sentimental expressions. We want to follow Zhang et al. [20] in analyzing the data with several models. From the outcomes of our experiments, we aim to gain insights into the predominant sentiments within the SWfMS domain. Additionally, we plan to implement our model on specific SWfMS platforms, conducting a comparative analysis to anticipate potential variations in user sentiments across these systems. Subsequently, we will adopt the approach proposed by Chang et al. [21] to extract critical factors influencing user satisfaction and SWfMS adoption. The analysis of these factors will provide us with a solid understanding of where to prioritize improvements, thereby ensuring enhanced user satisfaction. Our goal is to help the developers identify opportunities for the integration of sentiment analysis into development such as feature prioritization, bug identification, and user feedback. By adhering to this research methodology, developers can actively contribute to the development of SWfMSs that are more user-centric and operationally effective.

Anticipated Impact: Our research outcomes will boost researchers' confidence with positive sentiments in using software systems, particularly SWfMSs, and extend these concepts to other scientific fields. This empowers scientists to contribute more effectively to the scientific community. Additionally, our proposed solutions can enable software engineers to blend Software Engineering and Social Computing principles, enhancing interactions between scientists and software systems. We believe our study will be invaluable to scientists worldwide who rely on research-oriented software, optimizing their time and resources.

References:

[1] Q. Lu, X. Sun, Y. Long, Z. Gao, J. Feng, and T. Sun, "Sentiment Analysis: Comprehensive Reviews, Recent Advances, and Open Challenges," *IEEE Trans Neural Netw Learn Syst*, vol. PP, 2023, doi: 10.1109/TNNLS.2023.3294810.

[2] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P. M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," *Expert Syst Appl*, vol. 223, p. 119862, Aug. 2023, doi: 10.1016/J.ESWA.2023.119862.

[3] M. R. Islam and M. F. Zibran, "Leveraging Automated Sentiment Analysis in Software Engineering," *IEEE International Working Conference on Mining Software Repositories*, pp. 203–214, Jun. 2017, doi: 10.1109/MSR.2017.9.

[4] A. B. Eliacik and N. Erdogan, "Influential user weighted sentiment analysis on topic based microblogging community," Expert Syst. Appl., vol. 92, pp. 403–418, Feb. 2018.

[5] D. C. Edara, L. P. Vanukuri, V. Sistla, and V. K. K. Kolli, "Sentiment analysis and text categorization of cancer medical records with LSTM," *J Ambient Intell Humaniz Comput*, vol. 14, no. 5, pp. 5309–5325, May 2023, doi: 10.1007/S12652-019-01399-8/FIGURES/9.

[6] X. Pathak and M. Pathak-Shelat, "Sentiment analysis of virtual brand communities for effective tribal marketing," *Journal of Research in Interactive Marketing*, vol. 11, no. 1, pp. 16–38, 2017, doi: 10.1108/JRIM-09-2015-0069/FULL/XML.

[7] W. Deitrick and W. Hu, "Mutually Enhancing Community Detection and Sentiment Analysis on Twitter Networks," *Journal of Data Analysis and Information Processing*, vol. 2013, no. 03, pp. 19–29, Aug. 2013, doi: 10.4236/JDAIP.2013.13004.

[8] E. Choo, T. Yu, and M. Chi, "Detecting opinion spammer groups through community discovery and sentiment analysis," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9149, pp. 170–187, 2015, doi: 10.1007/978-3-319-20810-7_11/FIGURES/13.

[9] C. K. Pastor, "Sentiment Analysis of Filipinos and Effects of Extreme Community Quarantine Due to Coronavirus (COVID-19) Pandemic," *SSRN Electronic Journal*, Apr. 2020, doi: 10.2139/SSRN.3574385.

[10] T. G. Community *et al.*, "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update," *Nucleic Acids Res*, vol. 1, no. 1256879, pp. 13–14, 2013, doi: 10.1093/NAR/GKAC247.

[11] T. Oinn *et al.*, "Taverna: a tool for the composition and enactment of bioinformatics workflows," *Bioinformatics*, vol. 20, no. 17, pp. 3045–3054, Nov. 2004, doi: 10.1093/BIOINFORMATICS/BTH361.

[12] B. Ludäscher *et al.*, “Scientific workflow management and the Kepler system,” *Concurr Comput*, vol. 18, no. 10, pp. 1039–1065, Aug. 2006, doi: 10.1002/CPE.994.

[13] M. M. Hossain, B. Roy, C. K. Roy, and K. A. Schneider, “VizSciFlow: A Visually Guided Scripting Framework for Supporting Complex Scientific Data Analysis,” *Proc ACM Hum Comput Interact*, vol. 4, no. EICS, Jun. 2020, doi: 10.1145/3394976.

[14] J. Köster and S. Rahmann, “Snakemake—a scalable bioinformatics workflow engine,” *Bioinformatics*, vol. 28, no. 19, pp. 2520–2522, Oct. 2012, doi: 10.1093/BIOINFORMATICS/BTS480.

[15] R. Ferreira da Silva, R. Filgueira, I. Pietri, M. Jiang, R. Sakellariou, and E. Deelman, “A characterization of workflow management systems for extreme-scale applications,” *Future Generation Computer Systems*, vol. 75, pp. 228–238, Oct. 2017, doi: 10.1016/J.FUTURE.2017.02.026.

[16] T. Wollmann, H. Erfle, R. Eils, K. Rohr, and M. Gunkel, “Workflows for microscopy image analysis and cellular phenotyping,” *J Biotechnol*, vol. 261, pp. 70–75, Nov. 2017, doi: 10.1016/J.JBIOTEC.2017.07.019.

[17] S. Caíno-Lores, A. Lapin, J. Carretero, and P. Kropf, “Applying big data paradigms to a large scale scientific workflow: Lessons learned and future directions,” *Future Generation Computer Systems*, vol. 110, pp. 440–452, Sep. 2020, doi: 10.1016/J.FUTURE.2018.04.014.

[18] M. Kluge and C. C. Friedel, “Watchdog - a workflow management system for the distributed analysis of large-scale experimental data,” *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–13, Mar. 2018, doi: 10.1186/S12859-018-2107-4/FIGURES/12.

[19] R. Hoda, “Socio-Technical Grounded Theory for Software Engineering,” IEEE Trans. Softw. Eng., vol. 48, no. 10, pp. 3808–3832, Oct. 2022.

[20] T. Zhang, B. Xu, F. Thung, S. A. Haryono, D. Lo, and L. Jiang, “Sentiment Analysis for Software Engineering: How Far Can Pre-trained Transformer Models Go?,” Proc. - 2020 IEEE Int. Conf. Softw. Maint. Evol. ICSME 2020, pp. 70–80, Sep. 2020.

[21] V. Chang, L. Liu, Q. Xu, T. Li, and C. H. Hsu, “An improved model for sentiment a-nalysis on luxury hotel review,” Expert Syst., vol. 40, no. 2, p. e12580, Feb. 2023