Reflection

Zuxing Wu, a1816653

1 What was your research project?

My research project was designed to systematically identify and compare emerging technical trends in robotics and artificial intelligence (AI) by leveraging large-scale data from multiple online platforms. The motivation behind this work was to bridge the gap between different communities—developers, researchers, and practitioners—by analysing the diverse discussions and content found on **GitHub**, **Reddit**, **Stack Overflow**, and **arXiv**. Each of these platforms represents a unique facet of the robotics and AI ecosystem: GitHub for open-source development, Reddit for community-driven discussions, Stack Overflow for technical problem-solving, and arXiv for academic research.

To achieve a comprehensive and unbiased perspective, I developed a cross-platform analysis pipeline. This began with the automated collection of textual data using platform-specific APIs, which required careful planning to handle authentication, rate limits, and data structure differences. Once the data was gathered, I implemented rigorous data cleaning and preprocessing steps to remove noise, standardise formats, and ensure consistency across sources. This included removing HTML tags, filtering out irrelevant content, and normalising terminology.

The core of my analysis involved applying automated keyword extraction techniques, specifically using the YAKE algorithm, to identify the most salient technical terms and topics within each dataset. YAKE is an unsupervised, lightweight keyword extraction method that leverages local statistical features from single documents without requiring training data, and has demonstrated strong performance across diverse datasets and languages [2]. By aggregating and comparing the frequency and distribution of these keywords, I was able to uncover both shared and unique trends across platforms. The results were visualised using word clouds and comparative tables, which helped to highlight converging themes—such as the rise of large language models and human-robot interaction—as well as platform-specific focuses like implementation details or ethical considerations.

Ultimately, this project not only revealed the current technical directions shaping robotics and AI but also demonstrated the value of integrating heterogeneous data sources for a more holistic understanding of technological innovation. The findings provide actionable insights for researchers, developers, and policymakers interested in the evolving landscape of robotics and AI, and underscore the importance of interdisciplinary collaboration in addressing complex, real-world challenges.

2 What did you learn about doing research as part of your project that you didn't know before?

Through this project, I learned the importance of data cleaning and preprocessing in ensuring the quality and reliability of research findings. Initially, I underestimated how much effort would be required to standardise and prepare data from different platforms, each with its own structure, noise, and inconsistencies. Addressing these challenges taught me to be meticulous and patient, as even small oversights in cleaning could significantly affect the results. For example, prior research has shown that developer activity on Stack Overflow—particularly answering questions—is positively associated with increased productivity and code contributions on GitHub, highlighting the interconnectedness of these platforms and the value of cross-platform analysis [5].

I also realised how crucial it is to select appropriate tools and methods for data collection and analysis, such as choosing the right APIs and keyword extraction algorithms. Navigating API limitations, rate limits, and authentication requirements was more complex than I expected, and I had to adapt my approach several times to ensure robust data retrieval. This process deepened my technical skills and gave me a better appreciation for the practical aspects of large-scale data gathering.

Additionally, I gained a deeper appreciation for the challenges of integrating and comparing heterogeneous data from multiple sources. Each platform reflected different community norms and technical vocabularies, making direct comparison non-trivial. I discovered the value of visualisations in making complex results more interpretable, both for myself and for communicating findings to others.

On a personal level, this project helped me develop greater resilience and adaptability. I learnt to embrace setbacks as learning opportunities and to iterate on my methods rather than expecting perfect results on the first try. Overall, the experience broadened my understanding of empirical research and strengthened my confidence in tackling open-ended, interdisciplinary problems.

3 What would you do differently next time?

If I were to repeat this project, I would expand the range of platforms to include more diverse sources, such as Twitter, specialised robotics forums, or even industry blogs, to capture a broader and more nuanced spectrum of discussions and emerging trends. This would help ensure that the analysis reflects not only the perspectives of developers and researchers but also those of practitioners, enthusiasts, and the general public. I would also consider implementing more thoughtful data filtering, screening and selection criteria to ensure that the data collected is relevant and representative like Akpan et al. [1] did in their analysis.

I would also consider applying more advanced natural language processing techniques, such as topic modelling (e.g., LDA) or contextual embeddings (e.g., BERT), to uncover deeper semantic relationships and latent themes within the data. These methods could provide richer insights into the context and evolution of technical discussions, going beyond surface-level keyword frequency.

Furthermore, I would allocate more time to iterative testing and validation of the data extraction and analysis pipeline. This would involve conducting pilot studies, refining data cleaning procedures, and systematically checking for errors or inconsistencies at each stage. Establishing a more robust validation framework would improve the reproducibility and reliability of the results.

I noticed that recent research, such as the DancingLines framework, has demonstrated the value of combining semantic quantification, advanced time series alignment (e.g., vDTW-CD), and neural forecasting models like LSTM for cross-platform event popularity prediction. Incorporating similar multi-faceted analytical approaches in future work could enhance the accuracy and depth of trend analysis across diverse data sources [3].

Another improvement would be to incorporate feedback from domain experts or stakeholders throughout the research process. Engaging with experts could help refine the selection of keywords, interpret ambiguous findings, and ensure that the analysis remains relevant to real-world challenges in robotics and AI.

Finally, I would enhance the documentation and automation of the workflow, making it easier for others to reproduce or extend the study. This could include providing detailed code comments, sharing data processing scripts, and creating clear guidelines for adapting the pipeline to new platforms or research questions. By making these adjustments, future iterations of the project would be more comprehensive, rigorous, and impactful.

4 What is your advice to someone who is going to work on a similar project?

My main advice is to approach the project with careful planning and flexibility. Start by clearly defining your research objectives and the scope of your analysis, as this will guide your decisions on platform selection, data collection methods, and analytical techniques. When working with data from multiple sources, expect significant variation in data formats, terminology, and quality—so allocate ample time for data cleaning and preprocessing. This step is often more time-consuming than anticipated but is crucial for ensuring reliable results.

Familiarise yourself with the APIs and data access policies of each platform early on, and be prepared to handle issues such as authentication, rate limits, and incomplete or inconsistent data. Document your workflow thoroughly, including your data cleaning steps, code, and parameter choices, to make your research reproducible and easier to revisit or share with others.

Don't hesitate to iterate on your methods. Pilot studies and small-scale tests can help you identify potential problems before you commit to large-scale data collection or analysis. Use visualisation tools to explore your data and communicate your findings—visual summaries like word clouds or comparative charts can make complex results more accessible. Creating tables that contains chronologically-ordered data grouped by organisations or authors can also help in understanding the technical trends[4].

Seek feedback from peers, domain experts, or potential users of your research throughout the process. Their insights can help you refine your approach, interpret ambiguous results, and ensure your work remains relevant to real-world challenges. Finally, be patient and persistent: interdisciplinary projects often involve unexpected setbacks, but each challenge is an opportunity to learn and improve your research skills. Embrace the iterative nature of the process, and remember that clear documentation and openness to feedback are key to producing impactful and credible research.

References

[1] Ikpe Justice Akpan, Yawo M. Kobara, Josiah Owolabi, Asuama A. Akpan, and Onyebuchi Felix Offodile. 2025. Conversational and generative artificial intelligence and human–chatbot interaction in education and research. *International transactions in operational research* 32, 3 (2025), 1251–1281.

[2] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information sciences* 509 (2020), 257–289.

[3] Xiaofeng Gao, Wenyi Xu, Zixuan Zhang, Yan Tang, and Guihai Chen. 2023. Cross-Platform Event Popularity Analysis via Dynamic Time Warping and Neural Prediction. *IEEE transactions on knowledge and data engineering* 35, 2 (2023), 1337–1350.

[4] Bojan Obrenovic, Xiao Gu, Guoyu Wang, Danijela Godinic, and Ilimdorjon Jakhongirov. 2025. Generative AI and human–robot interaction: implications and future agenda for business, society and ethics. *AI & society* 40, 2 (2025), 677–690.

[5] Bogdan Vasilescu, Vladimir Filkov, and Alexander Serebrenik. 2013. StackOverflow and GitHub: Associations between Software Development and Crowdsourced Knowledge. In *2013 International Conference on Social Computing*. 188–195. doi:10.1109/SocialCom.2013.35