The Impact of the Advancement of Natural Language Processing on Open-source Software Community: A Preliminary Study

Meng Ma, Haskayne School of Business, University of Calgary
Giovani Da Silveira, Haskayne School of Business, University of Calgary

Abstract

This paper presents a preliminary investigation into the impact of advancements in Natural Language Processing technology, particularly in the context of the Open-Source Software community. Leveraging both the knowledge-based theory and the social identity theory, we examine the potential disruptive effects of NLP integration within this influential sphere of the software industry. An exploratory conceptual map is designed, elucidating the possible causal pathways through which NLP technology could influence the dynamics of open-source community. To validate the proposed theoretical model, we suggest conducting a quantitative analysis using survey data, offering individual-level insights into the hypotheses embedded within our framework. Recognizing the intricacies involved in such an investigation, we engage in discussion of underlying assumptions and potential challenges. Depending on various possible experiment outcomes, we provide a discourse on both the research and practical implications, thus providing a roadmap for future research and strategic planning in an area of AI-driven software development.

INTRODUCTION

Artificial Intelligence (AI) is redefining the trajectory of the modern world in unprecedented ways. As a catalyst for global transformation, AI has transcended conventional limits, delivering solutions to complex problems, and augmenting human capabilities across various domains. It is not only automating repetitive tasks, but also revolutionizing sectors from healthcare to finance, fostering productivity and efficiency, and thus reshaping our socio-economic structures (Russell et al., 2015). Among the various tools within the AI paradigm, Natural Language Processing (NLP) holds a distinctive position. NLP stands at the nexus of linguistics and machine learning, enabling machines to understand, generate, and interact using human language (Chowdhary, 2020). It plays an instrumental role in bridging the communication gap between humans and machines. In a digital era where data is the new oil, NLP extracts insights from the massive trove of unstructured textual data, driving the discovery of novel patterns and trends (Eisenstein, 2019). The significance of NLP, therefore, extends beyond mere technological intrigue, touching upon the very essence of human-machine symbiosis and the future of intelligent systems.

However, the use of NLP is not without controversy, catalyzing vigorous debate within academia and industry alike. Central to this discourse is the ethical quandary surrounding its deployment. NLP, particularly in its advanced forms like transformer-based models, can generate human-like text, which raises concerns over its potential misuse in generating disinformation or 'deepfake' text, undermining public discourse and information veracity (Goldstein et al., 2023). Furthermore, biases inherent in training data can be learned and

perpetuated by NLP systems, leading to outputs that may discriminate or stereotype, unintentionally yet tangibly (Kirk et al., 2021). Privacy implications are another focal point, as NLP tools often operate on sensitive personal or corporate data (Hacker, Engel, & Mauer, 2023). These controversies underscore the complexity of NLP's societal impact, necessitating ongoing discussion and investigation.

In our quest to understand the broader societal impact of Natural Language Processing (NLP), we aim to examine its implications within the realm of the Open-Source Software (OSS) community - an avenue of investigation that embodies a unique and revealing perspective. “The purpose of open source and free software licensing is to permit and encourage the involvement by licensees in improvement, modification, and distribution of the licensed work” (Laurent, 2004, p.164). This communal approach to software development has been foundational to the progress of the digital age, catalyzing innovations from small scale applications to vast digital infrastructures (Bonaccorsi, & Rossi, 2003). Probing the influence of NLP on the OSS community can yield insightful reflections on the wider societal impact of AI technologies. The OSS community is characterized by its vibrant collaboration, diverse skill set, and a shared mission to create and maintain freely available software (Batarseh, Kumar, & Eisenberg, 2020). It serves as a microcosm of the larger society that thrives on knowledge sharing, mutual learning, and cooperative endeavor - elements intrinsic to any social fabric. The impending AI/NLP wave could perturb these intricate dynamics and affect the community's prosperity, making it a pertinent and meaningful subject of study. Through this investigation, we hope to better understand, anticipate, and address potential adverse effects of NLP adoption in the OSS community, thus contributing to the mindful adoption and deployment of AI technologies.

Prior authors have documented the game-changing power of NLP in various domains. In the medical field, NLP techniques have been leveraged for extracting information from patient narratives, enhancing the precision of disease diagnosis (Meystre & Haug, 2006). In the industrial sector, NLP has played a pivotal role in improving machine-human interaction, boosting productivity, and reducing errors (Mah, Skalna, & Muzam, 2022). Moreover, the financial sector has utilized NLP tools for fraud detection, sentiment analysis and prediction of market trends, which has facilitated decision-making processes (Fisher, Garnsey, & Hughes, 2016). Despite these transformative applications, most authors emphasize the macro-level implications of NLP. We, however, shift the focus towards the micro-level, specifically, to individuals within the OSS community. This perspective offers an understanding of the impacts of NLP at the human level within communal structures. Considering the OSS community as a knowledge-intensive, collaborative, and diverse society, our investigation could potentially illuminate profound and nuanced effects of NLP adoption. This, in turn, will inform strategies for managing technological transitions in a way that is mindful of the social fabric of these communities.

In the following sections of this paper, we delve deeper into the complex intersection of NLP, the OSS community, and underlying theories of knowledge-based and social identity. We commence by expounding upon the concepts of NLP and OSS, their status, and core values, setting the stage for a comprehensive understanding of the intricate dynamics within these domains. Next, we introduce the knowledge-based theory and social identity theory, underpinning our hypotheses and theoretical rationale. The following hypotheses section presents a conceptual map that knits these theories together, proposing eight hypotheses that probe the multifaceted effects of NLP adoption within the OSS community.

Subsequently, the operationalization section unveils the fieldwork plan and how we intend to gather and analyze data, highlighting potential measures and biases that could emerge. Lastly, we conclude with a thought-provoking discussion and conclusion section, where we ponder over the expected results, their implications, and the potential mitigation solutions.

BACKGROUND

Artificial Intelligence and Natural Language Processing

Artificial Intelligence (AI) and Natural Language Processing (NLP) have both emerged as pivotal technologies in the information era. AI, in its essence, is a branch of computer science that mimics human intelligence in machines, enabling them to learn, reason, and problem-solve (Winston, 1984). It has its roots in the mid-20th century, with the first AI programs being written in the 1950s and 60s, designed to mimic human problem-solving and learning abilities. Over the decades, AI has grown to encompass a multitude of complex applications in areas such as autonomous vehicles, facial recognition, and predictive analytics, supporting a transformative shift in how we interact with technology (Zhang & Lu, 2021).

As a subset of AI, Natural Language Processing (NLP) focuses on the interface between computers and human language, providing machines with the capability to understand, interpret, and generate human language (Chowdhary, 2020). The development of NLP can also be traced back to the 1950s, with early efforts such as machine translation projects (Weaver, 1955). However, it was not until the advent of statistical methods and machine learning in the 1980s and 90s that NLP began to evolve into its contemporary form (Brown et al., 1990).

Today, NLP has a broad range of applications that continue to expand across numerous fields. These applications include sentiment analysis in market research (Berger et al., 2020), automatic text summarization in media (Liu & Lapata, 2019), speech recognition and synthesis in communication technology (Kamath, Liu, & Whitaker, 2019), and clinical decision support systems in healthcare (Fairie et al., 2021). As such, the impact of AI and NLP in shaping our world is profound and continues to grow, underlining the importance of investigating their societal implications.

Open-source Software and the Community

Open-source software (OSS) signifies a method of software development that promotes access to the end product's source materials. The OSS movement came to the fore in the late 1990s as a response to proprietary software models, which prohibited users from modifying or sharing the source code (Batarseh, Kumar, & Eisenberg, 2020). The principle behind OSS is that by freely sharing the code, the software is continually refined and enhanced by the collective contribution of programmers who improve upon it.

The OSS community is a collective of individuals and organizations committed to this collaborative ethos. These communities, which range from small groups focusing on specific projects to large networks like GitHub and SourceForge, comprise contributors who range from hobbyists to professionals. Over time, these communities have become important platforms for innovation, knowledge sharing, and learning, attracting a wide variety of developers with diverse skills and experiences (Feller & Fitzgerald, 2002).

The core values of the OSS community are rooted in notions of cooperation, transparency, and collective intelligence. Central to this is the belief that open collaboration can lead to more robust, innovative, and adaptable software than a closed, proprietary model. The success of this model is demonstrated by the wide adoption of OSS products in various sectors, from operating systems like Linux (Tu, 2000) and Android (Ableson, Sen, & Collins, 2009) to web servers like Apache (Mockus, Fielding, & Herbsleb, 2000).

Moreover, the OSS community plays a critical role in training and career development for many developers. According to Attwell (2005), it offers a platform for individuals to hone their skills, learn from others, and gain recognition for their contributions. The vitality of the OSS community thus reflects not just the success of a software development model, but also a significant social phenomenon in the digital age.

Knowledge-based Theory

One tenet in Knowledge-Based Theory (KBT), originating from the broader umbrella of the resource-based view of the firm, is that knowledge is the most strategically valuable resource of a firm or organization (Grant, 1996). Developers of this theory, which evolved in the late 1990s, underscore the pivotal role of knowledge and learning in propelling economic performance (Sveiby, 2001).

A central theme in KBT is the importance of knowledge sharing within an organization or community. Various researchers have significantly contributed to this facet of KBT. Notably, Grant (1996) viewed the firm as an institution for integrating knowledge, which inherently necessitates effective knowledge sharing. Moreover, Spender (1998) emphasized the role of managerial cognition and social complexity, hinting at the social dynamics that facilitate or inhibit knowledge sharing.

Kim and Nelson's (2000) work has been particularly instrumental in exploring the facilitators of knowledge sharing. They posited that a firm's ability to create and transfer knowledge internally is a primary source of competitive advantage. This perspective brings attention to factors like trust, shared language and codes, and an organizational culture that encourages sharing and collaboration.

In business research, KBT has informed studies on organizational learning, innovation, and dynamic capabilities, all of which are linked to knowledge sharing (Gurteen, 1999; Ipe, 2003; Jones, Cline, & Ryan, 2006). From a practical standpoint, KBT implies that managerial decisions should foster an environment conducive to knowledge sharing. This can involve nurturing a cooperative organizational culture, implementing systems to facilitate knowledge exchange, and recognizing and rewarding knowledge sharing behaviors (Gurteen, 1999).

Therefore, KBT is a seminal perspective in strategic management, stressing that a firm's success is contingent not just on what firms do, but also on what they know and how effectively they share that knowledge, as suggested by Small and Sage (2005). This highlights the necessity of focusing on the social and organizational mechanisms that enable knowledge sharing in a firm or community.

Social Identity Theory

Social Identity Theory (SIT), a psychological framework by the work of Tajfel and Turner (2004), offers profound insights into group behavior, intergroup relations, and the formation of group identities. This theory suggests that people's sense of self is shaped

significantly by the groups they identify with, leading to an internalized group membership that influences behavior.

The development of SIT owes much to numerous scholars who expanded on Tajfel and Turner's (2004) initial concepts. Notably, researchers have delved into the nuances of in-group bias, social categorization, and the interplay of personal and social identities (Hornsey, 2008). They found that group membership served as a source of pride and self-esteem, leading individuals to favor their in-group over out-groups, a behavior known as in-group bias (McLeod, 2008).

In the business research realm, SIT has been utilized to understand organizational behavior, leadership, team dynamics, and employee motivation (Ashforth, & Mael, 1989). For example, it has been used to explain why employees are more motivated when they feel a keen sense of belonging to their workgroup and organization (van Knippenberg, 2000). The theory also elucidates the phenomenon of organizational citizenship behavior, where employees go beyond their formal roles to contribute to the organization because they identify with it (Ellemers, de Gilder, & Haslam, 2004).

Furthermore, social identity can shape an organization's culture and reputation, as group identities often coalesce into a collective culture that influences the group's image both internally and externally (Moreland, Levine, & McMinn, 2001). It may also affect legitimacy, as the acceptance and validation of a group's identity by external stakeholders can significantly impact an organization's legitimacy (King, & Whetten, 2008; Spears et al., 2010).

Social Identity Theory furnishes a valuable lens for comprehending individual and group behaviors within organizations, emphasizing the vital role of social identity in shaping motivation, culture, reputation, and legitimacy. As such, the importance of fostering positive social identities should not be underestimated in the effective management of organizations and communities.

HYPOTHESES DEVELOPMENT

Conceptual Map

Based on Knowledge-Based Theory (KBT) and the Social Identity Theory (SIT), we propose a conceptual map (Figure 1) that indicates potential adverse effects of NLP on the OSS community. In the upper thread of the map, stemming from KBT, we posit that the progression of NLP technology facilitates an increased prevalence of NLP-powered bots. The deployment of these bots can inadvertently diminish human interactions and consequently attenuate the pivotal process of knowledge sharing among community members (Ferrara, 2016).

Simultaneously, advances in NLP, as depicted in the lower thread of the map inspired by SIT, have the potential to disrupt the balance of a diverse demographic in terms of skills and experiences within the community. This demographic imbalance could then negatively impact the communal sense of identity, an essential facet of social cohesion and individual fulfillment (Moharil et al., 2022).

Crucially, both knowledge sharing (Sowe, Stamelos & Angelis, 2008) and the sense of identity (Shih, & Huang, 2014) can be integral to the prosperity of the OSS community. Knowledge sharing may foster an environment of collaborative innovation, and a robust sense of identity may reinforce group cohesion and commitment. The potential erosion of

these crucial elements due to the proliferation of NLP-powered bots serves as the core concern driving this conceptual map.

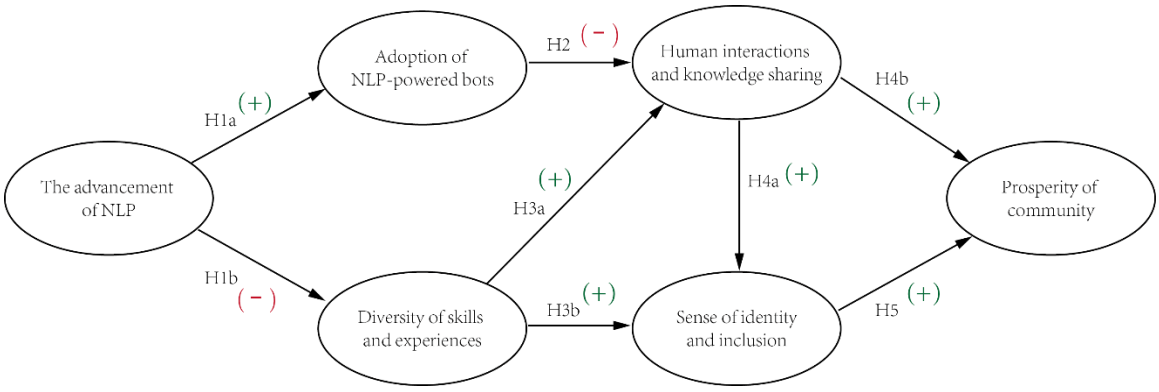

Figure 1. Conceptual map of the proposed theory

Research Hypotheses

Hypothesis 1a aligns with a core pursuit of the software development communities: leveraging technology to improve efficiency and effectiveness. The OSS community has exhibited consistent interest in embracing emerging technologies in its history (Bonaccorsi & Rossi, 2003). As NLP technology advances, the performance of NLP-powered bots, in terms of their understanding, generation, and interaction capabilities, also improves. Consequently, these enhanced bots are increasingly appealing for deployment on open-source platforms due to their potential to automate routine tasks, provide real-time support, and even contribute to code development (Svyatkovskiy et al., 2020). Moreover, the deployment of these bots can be highly cost-efficient. They can operate around the clock without the need for breaks and are not susceptible to human errors arising from fatigue or lapses in concentration. This consistent, efficient operation can facilitate faster project completion, thereby providing further incentive for their adoption. Hence, it is reasonable to hypothesize that:

Hypothesis 1a: The better the capability of NLP technologies, the greater the deployment of NLP-powered bots on open-source platforms.

Similarly, the influx of NLP technologies into OSS platforms presents another dynamic with respect to the diversity of skills and experiences among community members. As NLP technologies evolve, they have the potential to automate several tasks traditionally performed by human programmers (George & George, 2023). This is particularly relevant to junior-level programmers whose tasks often consist of routine and repetitive coding activities that are prime candidates for automation. Consequently, demand for these junior-level programmers may decline, reducing the representation of less-experienced members in the community. On the other hand, as the productivity of AI-powered human programmers increases due to NLP advancements, the community may see an increased representation of these highly skilled members. This potentially exacerbates the imbalance by over-representing certain skillsets, particularly those associated with AI and machine learning. Furthermore, NLP advancements may inadvertently promote the homogenization of skillsets. As more tasks become automated, programmers may need to adapt by focusing

on more complex, less routine tasks, thereby converging their skills towards similar areas (Peng et al., 2023). Given these reasons, it is hypothesized that:

Hypothesis 1b: The better the capability of NLP technologies, the lower the diversity of skills and experiences within community members.

As we delve deeper into the implications of technology on the human fabric of the OSS community, it becomes crucial to evaluate the effects of widespread bot adoption. Bots, despite their efficiency and consistency, cannot replicate the nuanced, multi-dimensional nature of human communication (Wessel et al., 2022). Interactions between community members are more than mere exchanges of information; they involve shared understanding, empathy, encouragement, and often, serendipitous learning. These elements, intrinsic to human interactions, are absent in bot-mediated communication. Undeniably, the increased presence of bots can alter the dynamics of access to human users. As bots become more prevalent, the chance of having meaningful conversations with human users might diminish. Human users, particularly newcomers, could find themselves interacting more with bots than with other humans (Ferrara, 2022). Additionally, knowledge sharing in OSS communities often thrives through learning by teaching and experience sharing. These rich forms of learning may be compromised with the increasing reliance on bots. While bots can provide helpful information, they are less adept at providing experiential insights or context-specific advice derived from personal experiences (Wessel et al., 2023). This leads to the next hypothesis:

Hypothesis 2: The greater the adoption of NLP-powered bots on open-source platforms, the less frequent the human interaction and knowledge sharing on these platforms.

The sheer spectrum of knowledge in a diverse community enhances the richness of knowledge sharing. Members bring unique skills and experiences to the table, increasing the breadth and depth of knowledge available to others. Consequently, a wider variety of problems can be solved, and more innovative solutions can be generated (Bell et al., 2018). Additionally, diversity engenders a plurality of perspectives that can enhance the quality of knowledge sharing. Different approaches to problem-solving can spark robust discussions and exchanges, leading to a more comprehensive understanding of the subject at hand. This cross-pollination of ideas is a hallmark of effective knowledge sharing (Salas, Reyes, & McDaniel, 2018). An inclusive environment, often found in diverse communities like OSS community, also fosters collaboration and encourages members to share knowledge freely. Inclusive cultures emphasize respect for individual contributions, thereby promoting an atmosphere conducive to knowledge sharing (Bodla et al., 2018). Lastly, mentorship is another key element in knowledge sharing (Bencsik, Juhász, & Machova, 2014). Senior members with diverse backgrounds can mentor less experienced members, sharing their insights and experiences, thereby promoting a culture of learning and knowledge sharing. Thus, it can be proposed that:

Hypothesis 3a: The higher the diversity of skills and experiences among members of the OSS community, the higher the frequency and quality of knowledge sharing in the community.

Role identity is a vital social identification aspect in OSS communities. As members obtain alternative skills and experiences, they can each carve out a unique role within the community, which can contribute to a stronger sense of identity (Gwebu, & Wang, 2011). Secondly, the sense of belonging, a key facet of social identity, can be bolstered by diversity. When members see their unique skills and experiences valued and utilized within the

community, they are more likely to feel a sense of belonging as indicated by Ye and Kishida (2003). Thirdly, the legitimacy of one's position within the community can also be enhanced by diversity. With a wide range of skills and experiences, members can demonstrate their value to the community in many ways, thereby validating their position. Additionally, diversity promotes the formation of groups within the community, each with its unique focus, skills, and goals. These groups can foster a secondary level of social identity as members identify with their group and the shared objectives within that group (Hertel, Niedner, & Herrmann, 2003). Further, common goals among group members, especially when these goals align with individual skills and experiences, can strengthen both group level identity as well as community level, especially so when, in diverse communities, these goals can be broad and inclusive, accommodating a wide range of skills and experiences, and thereby reinforcing the sense of inclusion (Ke & Zhang, 2009). Therefore, it can be proposed that:

Hypothesis 3b: The higher the diversity of skills and experiences among members of the OSS community, the stronger the sense of social identity and inclusion.

Naturally, the association between human interactions and social identity can be prevalent and profound. Human interactions through knowledge sharing can lead to individual empowerment and confidence within the community. As community members actively participate in knowledge sharing, they can foster a sense of importance and influence in the community, leading to a heightened sense of social identity (Ye & Kishida, 2003). Similarly, sharing experiences and expertise often creates common goals and bonds among community members. Shared experiences can also lead to a mutual understanding and a common narrative, thus reinforcing the community's identity and fostering a sense of inclusion. Further, the quality of knowledge sharing can directly affect the reputation and status of individuals within the OSS community. High-quality contributions increase the respect and admiration of peers, bolstering the contributor's reputation and reinforcing their perceived status (von Krogh, 2012). Finally, active participation in knowledge sharing can foster a sense of legitimacy and acceptance in the community. As individuals contribute valuable knowledge, they are recognized as legitimate and prominent members of the community, enhancing their sense of social identity and inclusion (Spears et al., 2010). Based on these considerations, it is plausible to propose that:

Hypothesis 4a: The greater the frequency and quality of knowledge sharing in the OSS community, the stronger the sense of social identity and inclusion.

Knowledge sharing is a fundamental mechanism in the OSS community that fuels innovation and collaborative problem solving (Bonaccorsi, & Rossi, 2003). The more community members that share knowledge, the more collective intelligence the community can harness. This can accelerate the development process, improve the quality of solutions, and, in turn, result in more effective and innovative products. Likewise, the quality of knowledge shared has a direct impact on the community's output. High-quality knowledge contributions often embody expertise and deep understanding, which can elevate the overall knowledge base of the community (Chang, & Chuang, 2011). This contributes to the development of superior software and boosts the reputation of the community, attracting more contributors and users, which can be considered a sign of community prosperity. Moreover, an elevated level of knowledge sharing can also foster a healthy and vibrant community culture. It not only facilitates mutual learning and growth among members but also fosters a sense of trust, reciprocity, and camaraderie (Chiu, Hsu, & Wang,

2006). This vibrant culture can make the community a more attractive place for potential contributors, further driving its growth and prosperity. Based on these rationales, we propose that:

Hypothesis 4b: The greater the frequency and quality of knowledge sharing in the OSS community, the greater the prosperity of the community.

Finally, social identity theory suggests that individuals derive part of their self-concept from the groups they belong to, leading to enhanced commitment and participation in group activities (Ellemers, De Gilder, & Haslam, 2004). In the context of OSS communities, a keen sense of social identity and inclusion can stimulate an important level of engagement among members. This engagement drives the creation and development of software, the primary output and measure of prosperity in these communities (Choi et al., 2013). When community members feel a keen sense of identity and inclusion, they are more likely to contribute their skills and knowledge in a way that benefits the community. Their commitment and the resultant actions can fuel continuous innovation, foster an open, collaborative atmosphere, and attract more members with varied skills and knowledge - factors that contribute to community prosperity, as suggested by Shen, Yu, and Khalifa (2010). In addition, they also discovered that a keen sense of social identity and inclusion can cultivate a sense of loyalty and dedication, as this may decrease member attrition, ensuring the stability and sustainability of the community. It can also positively impact the reputation of the community, drawing further participation and usage, which are key indicators of community prosperity. Considering the crucial role social identity and inclusion play in motivating member participation and fostering community development, we propose that:

Hypothesis 5: The stronger the sense of social identity and inclusion within the OSS community, the greater the prosperity of the community.

OPERATIONALIZATION

The study population consists of participants in the OSS community. We will test our hypotheses with structural equation models using data from a stratified sample of OSS participants, including developers, project managers, event organizers, etc. The stratified sample aims not only to increase representation but also to minimize the impact of self-selection bias, as individuals in certain positions might be more willing to participate as their work might be impacted by NLP in specific ways.

We will collect data on the several factors in the research map including the degree of adoption of NLP-powered bots, diversity of skills and experiences among community members, and the quality and frequency of knowledge sharing within the community. We will also collect data on participants' perceptions of their social identity and sense of inclusion, as well as their perspectives on community prosperity. We will test the validity and reliability of the measurement model using well-known techniques in confirmatory factor analysis (Fornell and Larcker, 1981; MacKenzie; Podsakoff and Podsakoff, 2011).

In addition to survey data, we will supplement our findings with publicly available data from OSS platform metrics. This auxiliary data can provide additional insights and may serve as control variables or to support further reliability analyses through triangulation with the primary survey data.

DISCUSSION AND CONCLUSION

We investigate the implications of NLP technology adoption for the OSS community. Based on knowledge-based and social identity theories, we propose a conceptual map that links the progression of NLP to the dynamics within the OSS community, and the community's overall prosperity. We plan to test our theory using structural equation modeling of primary and secondary data from the OSS community.

We acknowledge certain assumptions that underpin the study. Firstly, our focus is on the current state of NLP technology, mainly generative language models, rather than the more advanced Artificial General Intelligence (AGI). While AGI developments are underway, most current applications of AI in software development communities involve NLP tools that fall in the category of narrow AI rather than full-fledged AGI (Ng & Leung, 2020). Secondly, in this study we do not consider the impact of corporate funding on the OSS community, despite its relevant impact in community prosperity (Vetter, 2021). Lastly, our data collection will rely heavily on perception and recall of survey respondents, without prior specification of lags associated with the time effects of predictors on outcome variables. Short-term effects might include an initial boom in the OSS community with the integration of advanced NLP tools, while long-term effects may be still largely unknown.

Study results may yield a variety of outcomes, each with its own implications for our understanding of NLP's influence on the dynamics of OSS communities.

If our findings conform to knowledge-based theory, they will highlight the indelible role of human-to-human interaction, even amidst the growing capabilities of AI technologies. We might suggest that, despite the technological advancements, there exists an inherent value in human interactions that current AI may not replicate or replace. We could attribute this to several reasons, among which includes the fact that human interactions, often characterized by nuances and subtleties, are a vital part of knowledge sharing and collaborative problem solving in OSS communities (Seering et al., 2019). Moreover, alignment with knowledge-based theory would point towards the significance of human trial and error in contributing to progress within communities. This process is not just about finding errors, but also about learning, adapting, and understanding - an area where AI technologies still lag human capabilities. It underscores the essence of a learning culture, where mistakes are viewed as opportunities for growth and betterment, which is an intrinsically human concept. Furthermore, if the significance of human interaction is validated, it would indicate the importance of collective creativity in the realm of communal software development, an area where AI, despite its growing sophistication, continues to struggle. Similar collective innovation has been recently observed in the Python development community (Pike, 2022). These findings align with the belief that while AI can optimize and automate certain aspects of programming, innovative problem-solving capabilities are still unique to humans.

Conversely, if our results align with social identity theory, they will underscore the value that individuals attach to their role within the OSS community. We might suggest that simply belonging to a group is not solely about making contributions, but also about their moral worth to fit in the group. This would suggest that career progression and the opportunity to take on increasing responsibility carries intrinsic value. This signifies the importance of the development process (Csikszentmihalyi, 2013), asserting that the journey towards achieving the end goal may sometimes be more rewarding than the

achievement itself. Further, if the social identity aspect of theory is validated, it would point towards the importance of the collaborative process within the OSS community. Beyond the utilitarian value of collaboration, the act of working together could bring hedonic value, renewing joy, and motivation for individuals involved in their work and career.

Still, if the findings do not support either explanation, we may need to rethink the prevailing theoretical frameworks within OSS community research. This could suggest the presence of other influential factors that may have a larger role to play in the context of NLP technology integration. Such results may encourage us to broaden our model to include other dynamics within OSS communities such as technological adaptability, leadership structures, external funding, or regulatory environments. It also underscores the importance of using a multi-faceted approach in investigating OSS communities in the face of advancing NLP technology, hinting at the need for practitioners to look beyond the lens of the communal interaction to ensure the prosperity and survival of their communities.

In conclusion, we aim with this study to shed light on the complex dynamics between NLP technology adoption, knowledge sharing, social identity, and the prosperity of OSS communities. While acknowledging the limits of our assumptions and potential biases, we anticipate our findings can offer valuable insights and practical recommendations for managing OSS communities in an increasingly AI-infused landscape.

References

Ableson, F., Sen, R., & Collins, C. E. 2009. ***Unlocking Android***. Manning Publications.

Ashforth, B. E., & Mael, F. 1989. Social identity theory and the organization. ***Academy of Management Review***, 14(1): 20-39.

Attwell, G. 2005, July. What is the significance of Open Source Software for the education and training community. ***Proceedings of the First International Conference on Open Source Systems***: 353-358.

Batarseh, F. A., Kumar, A., & Eisenberg, S. 2020. The history and future prospects of open data and open source software. ***Data Democracy***: 29-43. Academic Press.

Bencsik, A., Juhász, T., & Machova, R. 2014. Mentoring practice on behalf of knowledge sharing in the light of education. ***Acta Polytechnica Hungarica***, 11(9): 95-114.

Bell, S. T., Brown, S. G., Colaneri, A., & Outland, N. 2018. Team composition and the ABCs of teamwork. ***American Psychologist***, 73(4): 349.

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. 2020. Uniting the tribes: Using text for marketing insight. ***Journal of Marketing***, 84(1):1-25.

Bodla, A. A., Tang, N., Jiang, W., & Tian, L. 2018. Diversity and creativity in cross-national teams: The role of team knowledge sharing and inclusive climate. ***Journal of Management & Organization***, 24(5): 711-729.

Bonaccorsi, A., & Rossi, C. (2003). Why open source software can succeed. ***Research policy***, 32(7): 1243-1258.

Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., et al. 1990. A statistical approach to machine translation**,** ***Comput Linguist***, 16 (2).

Chang, H. H., & Chuang, S. S. 2011. Social capital and individual motivations on knowledge sharing: Participant involvement as a moderator. ***Information & Management***, 48(1): 9-18.

Chiu, C. M., Hsu, M. H., & Wang, E. T. 2006. Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. Decision support systems, 42(3): 1872-1888. Eisenstein, J. 2019. ***Introduction to Natural Language Processing. MIT Press.***

Choi, J., Choi, J., Lee, H. S., Hwangbo, H., Lee, I., & Kim, J. 2013. The reinforcing mechanism of sustaining participations in open source software developers: based on social identity theory and organizational citizenship behavior theory. ***Asia Pacific Journal of Information Systems***, 23(3): 1-23.

Chowdhary, K. R. 2020. Natural language processing. ***Fundamentals of Artificial Intelligence***: 603-649.

Csikszentmihalyi, M. 2013. Flow: The psychology of happiness. ***Random House***.

Ellemers, N., De Gilder, D., & Haslam, S. A. 2004. Motivating individuals and groups at work: A social identity perspective on leadership and group performance. ***Academy of Management Review***, 29(3): 459-478.

Fairie, P., Zhang, Z., D'Souza, A. G., Walsh, T., Quan, H., & Santana, M. J. 2021. Categorising patient concerns using natural language processing techniques. ***BMJ Health & Care Informatics***, 28(1).

Feller, J., & Fitzgerald, B. 2002. ***Understanding Open Source Software Development***. Addison-Wesley Longman Publishing Co., Inc.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. 2016. The rise of social bots. ***Communications of the ACM***, 59(7): 96-104.

Ferrara, E. 2022. Twitter Spam and False Accounts Prevalence, Detection and Characterization: A Survey. *arXiv preprint arXiv*:2211.05913.

Fisher, I. E., Garnsey, M. R., & Hughes, M. E. 2016. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. Intelligent systems in accounting, ***Finance and Management*****,** 23(3):157-214.

Fornell, C., and D. F. Larcker. 1981. Evaluating structural equation models with unobservable variables and measurement error. ***Journal of Marketing Research****,* 18(1): 39-50.

George, A. S., & George, A. H. (2023). A review of ChatGPT AI's impact on several business sectors. ***Partners Universal International Innovation Journal***, 1(1): 9-23.

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. 2023. Generative language models and automated influence operations: emerging threats and potential mitigations. *arXiv preprint arXiv*:2301.04246.

Grant, R. M. 1996. Toward a knowledge-based theory of the firm. ***Strategic Management Journal***, 17(S2): 109-122.

Gurteen, D. 1999. Creating a knowledge sharing culture. ***Knowledge Management Magazine*****,** 2(5): 1-4.

Gwebu, K. L., & Wang, J. 2011. Adoption of Open Source Software: The role of social identification. ***Decision Support Systems***, 51(1): 220-229.

Hacker, P., Engel, A., & Mauer, M. 2023. Regulating ChatGPT and other large generative ai models. *arXiv preprint arXiv*:2302.02337.

Hertel, G., Niedner, S., & Herrmann, S. 2003. Motivation of software developers in Open Source projects: an Internet-based survey of contributors to the Linux kernel. ***Research Policy***, 32(7), 1159-1177.

Hornsey, M. J. 2008. Social identity theory and self-categorization theory: A historical review. ***Social and Personality Psychology Compass***, 2(1): 204-222.

Ipe, M. 2003. Knowledge sharing in organizations: A conceptual framework. ***Human Resource Development Review***, 2(4): 337-359.

Jones, M. C., Cline, M., & Ryan, S. 2006. Exploring knowledge sharing in ERP implementation: an organizational culture framework. ***Decision Support Systems***, 41(2): 411-434.

Kamath, U., Liu, J., & Whitaker, J. 2019. Deep learning for NLP and speech recognition Vol. 84. ***Cham, Switzerland: Springer***.

Ke, W., & Zhang, P. 2009. Motivations in open source software communities: The mediating role of effort intensity and goal commitment. ***International Journal of Electronic Commerce*****,** 13(4): 39-66.

Kim, L. and Nelson, R. R. 2000. Technology, learning, and innovation: Experiences of newly industrializing economies, ***Cambridge, UK: Cambridge University Press***.

King, B. G., & Whetten, D. A. 2008. Rethinking the relationship between reputation and legitimacy: A social actor conceptualization. ***Corporate Reputation Review***, 11: 192-207.

Kirk, H. R., Jun, Y., Volpin, F., Iqbal, H., Benussi, E., Dreyer, F., ... & Asano, Y. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. ***Advances in Neural Information Processing Systems*****,** 34: 2611-2624.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. 2011. Construct measurement and validation procedures in MIS and behavioral research: integrating new and existing techniques. ***MIS Quarterly***, 35(2):293–334.

McLeod, S. (2008). Social identity theory.

Mah, P. M., Skalna, I., & Muzam, J. 2022. Natural language processing and artificial intelligence for enterprise management in the era of industry 4.0. ***Applied Sciences***, 12(18): 9207.

Meystre, S., & Haug, P. J. 2006. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. ***Journal of Biomedical Informatics***, 39(6): 589-599.

Mockus, A., Fielding, R. T., & Herbsleb, J. 2000, June. A case study of open source software development: the Apache server. ***Proceedings of the 22nd International Conference on Software Engineering***: 263-272.

Moharil, A., Orlov, D., Jameel, S., Trouwen, T., Cassee, N., & Serebrenik, A. 2022. Between JIRA and GitHub: ASFBot and its influence on human comments in issue trackers. ***Proceedings of the 19th International Conference on Mining Software Repositories***: 112-116.

Moreland, R. L., Levine, J. M., & McMinn, J. G. 2001. Self-categorization and work group socialization. ***Social Identity Processes in Organizational Contexts***: 87-100.

Ng, G. W., & Leung, W. C. 2020. Strong artificial intelligence and consciousness. ***Journal of Artificial Intelligence and Consciousness***, 7(01): 63-72.

Laurent, A. M. S. 2004. ***Understanding open source and free software licensing: guide to navigating licensing issues in existing & new software.*** O'Reilly Media, Inc.

Liu, Y., & Lapata, M. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv*:1908.08345.

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. 2023. The impact of ai on developer productivity: Evidence from GitHub copilot. *arXiv preprint arXiv*:2302.06590.

Pike, T., Colter, R., Bailey, M., Kazil, J., & Meyers, J. S. 2022. Social Networks as a Collective Intelligence: An Examination of the Python Ecosystem. *arXiv preprint arXiv*:2201.06040.

Russell, S., Dewey, D., & Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence. ***AI Magazine***, 36(4): 105-114.

Salas, E., Reyes, D. L., & McDaniel, S. H. 2018. The science of teamwork: Progress, reflections, and the road ahead. ***American Psychologist***, 73(4): 593.

Seering, J., Luria, M., Kaufman, G., & Hammer, J. 2019, May. Beyond dyadic interactions: Considering chatbots as community members. ***Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems:*** 1-13.

Shen, K. N., Yu, A. Y., & Khalifa, M. 2010. Knowledge contribution in virtual communities: accounting for multiple dimensions of social presence through social identity. ***Behaviour & Information Technology***, 29(4): 337-348.

Shih, H. P., & Huang, E. 2014. Influences of Web interactivity and social identity and bonds on the quality of online discussion in a virtual community. ***Information Systems Frontiers***, 16: 627-641.

Small, C. T., & Sage, A. P. 2005. Knowledge management and knowledge sharing: A review. ***Information Knowledge Systems Management***, 5(3): 153-169.

Sowe, S. K., Stamelos, I., & Angelis, L. 2008. Understanding knowledge sharing activities in free/open source software projects: An empirical study. ***Journal of Systems and Software***, 81(3): 431-446.

Spears, R., Greenwood, R., de Lemus, S., & Sweetman, J. 2010. Legitimacy, social identity and power. ***The Social Psychology of Power***: 251-283.

Spender, J. C. 1998. Pluralist epistemology and the knowledge-based theory of the firm. ***Organization***, 5(2): 233-256.

Svyatkovskiy, A., Deng, S. K., Fu, S., & Sundaresan, N. 2020, November. Intellicode compose: Code generation using transformer. ***Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering***:1433-1443.

Sveiby, K. E. 2001. A knowledge-based theory of the firm to guide in strategy formulation. ***Journal of Intellectual Capital***, 2(4): 344-358.

Tajfel, H., & Turner, J. C. 2004. The social identity theory of intergroup behavior. ***Political Psychology***: 276-293. Psychology Press.

Tu, Q. 2000. Evolution in open source software: A case study. ***Proceedings 2000 International Conference on Software Maintenance***: 131-142. IEEE.

van Knippenberg, D. 2000. Work motivation and performance: A social identity perspective. ***Applied Psychology***, 49(3): 357-371.

Vetter, M. 2021. Acquisitions and open source software development. Springer Gabler.

von Krogh, G., Haefliger, S., Spaeth, S., & Wallin, M. W. 2012. Carrots and rainbows: Motivation and social practice in open source software development. ***MIS quarterly***: 649-676.

Weaver, W. 1955. Translation. Mach Transl Lang, 14, pp. 15-23 Wessel, M., Wiese, I., Steinmacher, I., & Gerosa, M. A. 2021. Don't disturb me: Challenges of interacting

with software bots on open source software projects. ***Proceedings of the ACM on Human-Computer Interaction***, 5(CSCW2): 1-21.

Wessel M, Serebrenik A, Wiese I, Steinmacher I and Gerosa M. 2022. Quality gatekeepers: investigating the effects of code review bots on pull request activities. ***Empirical Software Engineering***. 10.1007/s10664-022-10130-9. 27:5.

Wessel M, Zaidman A, Gerosa M and Steinmacher I. 2023. Guidelines for Developing Bots for GitHub. ***IEEE Software***. 40(3): 72-79.

Winston, P. H. 1984. ***Artificial Intelligence***. Addison-Wesley Longman Publishing Co., Inc.

Ye, Y., & Kishida, K. 2003. Toward an understanding of the motivation of open source software developers. ***25th International Conference on Software Engineering, 2003. Proceedings.***: 419-429. IEEE.

Zhang, C., & Lu, Y. 2021. Study on artificial intelligence: The state of the art and future prospects. ***Journal of Industrial Information Integration***, 23:100224.