Understanding Open-Source NLP Artifact Adoption Through Information Systems Success Factors

Kaiyue Liu
The University of Texas at Arlington
Kaiyue.Liu@uta.edu

Abstract

Open-Source Software (OSS) movement has significantly shaped the landscape of data science, particularly in the subfield of natural language process (NLP). Despite the popularity and rapid growth of OSS tools in the field of data science, prior IS literature did not examine the adoption of NLP artifacts through the lens of OSS success. This study applies and extends the DeLone and McLean Information Systems Success Model to the context of OSS in the domain of NLP. Our extended model examines the moderating effects of task type on the relationship between system quality, information quality, and adoption. In this study, we gather model cards of NLP artifacts, and their download/endorsement counts from Hugging Face and empirically examine the adoption behavior of OSS NLP artifacts. Our expected findings would suggest that system quality affects adoption more for analysis tasks compared to generation tasks, while information quality affects adoption more for generation tasks compared to analysis tasks.

Keywords: Open-Source Software, Information Systems Success, Natural Language Processing

1. Introduction

Data science, defined as "the study of the generalizable extraction of knowledge from data" according to Dhar (2013), has experienced rapid growth in recent years in both industry and academia. A search of the term "data science" on Google yielded 4.8 billion pages of results in mid-2023. The term on Google Trends has more than doubled in popularity from early-2016 to mid-2023, indicating its increasing and continuous growth. In the past decade, data science has become increasingly impactful and popular with the continuous advancement of both hardware and software (algorithms) capabilities, growth of data collection, and enhancement of business value. A key factor contributing to this success and growth is the democratization of data science, which has been largely enabled by the popularity and availability of open-source data science tools among both academia and practitioners (Kross et al., 2020).

Data science have a variety of subfields, each addressing specific challenges and tasks. As with other areas of data science, natural language processing (NLP) has also benefited from the availability and adoption enabled by OSS. With OSS facilitating cost-saving and creativity, contributing to the adoption of NLP artifacts, numerous NLP deep learning models for different tasks (such as text classification, question answering, summarization, text generation, etc.) were created, modified, and shared across the internet freely. On Hugging Face Hub, as of May 2023, there were 193, 810 deep learning models freely and openly distributed for all researchers and practitioners, and many of them were downloaded tens of millions of times each month. For example, the gpt2 model, which is a model for text generation, was downloaded over 23 million times in April 2023. Despite the complexity of these models and the difficulty of training them, OSS NLP tools such as Hugging Face enabled the general public to access state-of-the-art NLP technologies.

From an Information Systems (IS) perspective, the adoption of OSS, as one of the most important indicators of OSS success, has been a major topic that has attracted significant attention from scholars and researchers. Over the years, numerous studies have sought to understand the factors contributing to OSS success and the mechanisms through which these factors influence

URI: https://hdl.handle.net/10125/107311
978-0-9981331-7-1

adoption (e.g. Crowston et al., 2003; S.-Y. T. Lee et al., 2009; Gwebu and Wang, 2011; Rossi et al., 2012; Lenarduzzi et al., 2020). The open-source nature of these NLP tools played a crucial role in their adoption, allowing users and developers to create, modify and share models for their specific needs free of charge, contributing to their widespread adoption and success in both academia and industry. Despite the important role open-source philosophy played in the development of data science and NLP, there are not many studies in the field of IS studying the diffusion and adoption of NLP artifacts under the scope of open-source software. To address this literature gap, we propose this study to examine the adoption of NLP models through the lens of OSS success.

NLP task typology is also an interesting angle that has not been examined thoroughly by IS researchers. As different types of tasks share different characteristics, the driving factors for adoption may impact differently among different tasks. Although previous literature provided plenty of examples of artifacts in each type of task, no literature compares them, especially on how they affect adoption. Our study addresses this literature gap, by comparing numerous models with different types of primary tasks, we can examine the task typology and how it impacts the adoption: how does the adoption behavior differ between different types of tasks?

In this study, we aim to examine the success and adoption of NLP models through the lens of OSS success, applying and extending the DeLone and McLean Information Systems Success Model (DeLone and McLean, 2003). Our extended model reveals the moderating effects of task type on the relationship between system quality, information quality, and adoption. By focusing on NLP models within the context of data science and OSS, we seek to identify factors that contribute to their adoption and use in various applications. We position the research questions of this study as follows:

How does system/information/service quality affect the adoption and success of open-source NLP models? And how does NLP task type moderate the relationship between system/information/service quality and adoption?

Our research contributes to the literature in two key ways. We examine and extend the DeLone and McLean IS success model within the context of OSS success in Data Science. Our study aims to confirm the DeLone and McLean IS success model and its efficacy in explaining data science OSS adoption behaviors and extend the model to consider how types of tasks moderate this adoption behavior, theoretically expanding the original DeLone and McLean IS success model. By identifying the key factors driving NLP OSS adoption, this study sheds light on the design elements that contribute to the success of NLP models in the OSS ecosystem. By examining the empirical result and synthesizing design principles for OSS NLP models that facilitate better adoption, our research provides actionable insights for developers and researchers to create more successful and widely adopted NLP tools.

In the following section, we will first provide a comprehensive review of the existing literature on OSS success and the background of NLP models. We then propose a modified research model and develop the hypotheses for this study. The data collection, methodology used for analysis, and results will be presented in the method and result section.

2. Literature Review

2.1. OSS Data Science Tools and NLP Artifacts

Open-source software (OSS), according to Open Source Initiative, is software whose source code is openly distributed, allowing users to access, modify the software and redistribute the derived versions free of charge. OSS has been a phenomenon in the past 2 decades shaping the landscape of modern technology, as more and more back-office servers have been powered by OSS infrastructure applications, such as Apache, Linux, etc. (Fitzgerald and Kenny, 2004). With the rapid and diverse growth of OSS products along with the open-source philosophy, the open-source movement has taken over and become a common practice, especially in the field of computer science (Heron et al., 2013).

OSS provided many advantages compared to proprietary software. Not only is OSS free and can save costs for its users, but also the collaborative nature enabled rapid development and innovation that contributed to its success in the data science community (Vasilescu et al., 2015). Many successful data science tools are open-sourced, including database management tools, programming languages, and packages for specific uses, such as generating graphics, processing data, and NLP solutions for chatbots, etc. (Barlas et al., 2015). The OSS advantages such as easy access, encouraged more adoption in the field of education which further enhanced adoption overall. For example, Jupyter Notebook, an open-source software for interactive computing, has been widely adopted across many STEM and STEAM disciplines (Hanč et al., 2020).

Natural Language Processing (NLP), a field of study employs computers to understand and manipulate

natural language (Chowdhary and Chowdhary, 2020), has gained significant attention due to its potential to revolutionize human-computer interactions, enabling machines to understand, interpret, and generate human language in meaningful and useful ways (Ray et al., 2018; Manaris, 1998). In this rapid recent development, OSS also played an important role in the process. The openness and transparency provided by OSS NLP artifacts enabled researchers and developers to verify each other's work and build upon each other (Stodden et al., 2018). OSS NLP artifacts can be accessed freely, modified according to the developer's customized needs, and shared easily among the data science community via libraries such as spaCy and Hugginface (Wolf et al., 2020).

2.2. Hugging Face Transformers

In this study, we focus on the case of the Hugging Face Transformers library (Wolf et al., 2020). This library has gained significant importance in the NLP community. The success of the Hugging Face Transformers library can be attributed to the OSS philosophy, which promotes collaboration and sharing of resources among researchers and developers. Many artifacts have been developed using this library in both computer science and IS disciplines (e.g. Chen et al., 2020; Song et al., 2022).

The Hugging Face Transformers library serves as an excellent case for our study. The first advantage is its popularity. It is a major player in the NLP community with wide adoption already. Many popular models on the model-sharing platform, Hugging Face Hub, have monthly download counts in the tens of millions. This ensures that the findings from this study would be representative of the broader NLP community. The Hugging Face Transformers library also stands out for its inclusiveness and comprehensive coverage of NLP models. This extensive collection ensures that the data gathered from the library is representative and reflects the broad landscape of NLP models. By studying the adoption behaviors within the context of the Hugging Face library, we can improve our understanding of the adoption of various OSS NLP artifacts.

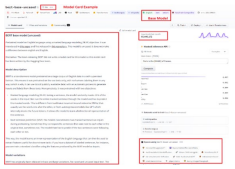

Figure 1. Example Hugging Face Model Card.

On top of the previously mentioned advantages, the data quality of model cards on Hugging Face Hub, as shown in the figure 1, is also significantly better than many other data sources, which facilitates our data gathering and analyzing processes (Mitchell et al., 2019). Hugging Face Transformers library is appropriate for us to investigate the adoption behaviors in NLP artifacts through the lens of OSS success.

2.3. Information Systems Success in OSS

Information systems success is one of the oldest and most used dependent variables in IS research, and in OSS literature, OSS success, in terms of quality (such as usability and maintainability), has been a major topic since the beginning of OSS (Ghapanchi et al., 2011). The pursuit of understanding IS success is critical as it helps researchers and practitioners to identify the factors that contribute to the effectiveness and value of information systems artifacts. By exploring the dimensions of IS success and developing models that capture more nuisances in the adoption behavior, our study can researchers to better understand the adoption behaviors, and facilitate practitioners on improving the adoption of OSS NLP artifacts.

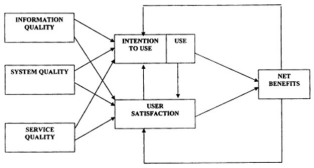

Figure 2. DeLone and McLean, 2003 Model.

The most cited IS success model in the context of OSS is the DeLone and McLean (2003) IS Success model, as shown in Figure 2. DeLone and McLean's model posits that quality has three major dimensions: system quality, information quality, and service quality. These dimensions together influence user behavior and satisfaction, ultimately contributing to the overall success of an open-source information system. In their 2003 update, DeLone and McLean expanded the model to encompass service quality, which in the context of OSS can be considered as community service quality. Since in OSS, the service is typically provided freely by other users in the form of user-to-user interactions (Lakhani and Von Hippel, 2004; S. M. Lee and Lee, 2012). The model is supported by the literature well that it has been frequently used in IS literature on OSS success (e.g., Alotaibi and Alshahrani, 2022). We adapt this model into our context to explain the adoption of OSS NLP artifacts.

In this study we deploy the DeLone and McLean (2003) IS success model because it is the most appropriate fit for our study. The application of the DeLone and McLean model in our study is firmly grounded in existing literature, where it has been

widely used in the OSS context (e.g. Crowston et al., 2006; Hollmann et al., 2013). Its focus on the use-environment, rather than the development environment, aligns perfectly with our investigation into NLP artifacts, where training and fine-tuning are standardized procedures (Subramaniam et al., 2009). Considering these specific attributes of the OSS NLP context, the concerns regarding biases or limitations inherent to the model is less relevant. Therefore, we conclude that the DeLone and McLean model serves as a suitable and robust theoretical framework for our study, allowing us to explore the complexities of adoption without significant concerns about bias.

The DeLone and McLean IS success, compared to other models, namely the Technology Acceptance Model (TAM) (Davis, 1989) and the Diffusion of Innovation theory (DOI) (Rogers et al., 2014), fits our OSS and NLP contexts much better. Not only DeLone and McLean's IS success model is well supported by prior literature (e.g. Subramaniam et al., 2009; Howison and Crowston, 2014), but also it is the most suitable framework among competing theories of TAM and DOI. While TAM offers valuable insights into user adoption from the perspective of perceived ease of use and usefulness, it falls short in addressing critical technical complexities intrinsic to AI NLP artifacts. These complexities include the type of model, its cost, performance metrics, and broader community factors that are paramount in the NLP context. DOI, despite its broader scope, is less apt for our research due to its general nature that incorporates a myriad of variables (e.g. innovation, task, individual characteristics, environmental conditions, and organizational structure). This broad focus poses challenges in operationalizing individual and organizational constructs and could dilute the emphasis on the importance of community service quality in OSS adoption patterns. Therefore, DeLone and McLean's model offers the most balanced approach by accommodating both user-centric and technical aspects, thus making it the most apt theoretical framework for our study.

2.3.1. System Use and User Satisfaction Defining IS success can be challenging due to its subjective and multi-dimensional nature, particularly for OSS projects where determining the intended user base and expected outcomes can be even more difficult. Due to these difficulties, many studies employed systems use as the indicator of information system success, especially in the context of OSS (Crowston et al., 2003).

System use can be quantified through various metrics such as download counts, active users, and frequency of use, which are easier to measure and provide a tangible perspective on the success of an OSS project. Prior IS studies suggest popularity, measured by download counts is a primary metric for adoption, as popularity is central to the health and functionality of OSS projects (e.g. Subramaniam et al., 2009; Howison and Crowston, 2014). Though this metric may not directly capture all facets of adoption depth or quality, it is a widely accepted and meaningful predictor in the literature. This approach aligns with the specific nature of OSS NLP artifacts, where usage typically involves direct downloading and deployment without extra developmental stages. Thus, the relevance of depth and quality of adoption is less pronounced.

Another important dependent variable in the context of our research is user satisfaction. User satisfaction can also be operationalized through surveys, reviews, endorsements, comments, and feedback (e.g. S. M. Lee and Lee, 2012; Borges et al., 2016; Blincoe et al., 2016). This operationalization advantage enables our study using field data to investigate the other important use construct in the DeLone and McLean model, user satisfaction. By combining both system use and user satisfaction, our findings can extend the understanding and evaluation of the OSS success, especially for OSS NLP artifacts.

2.3.2. System, Information, and Service Quality The quality of OSS projects includes three main dimensions, Information Quality, System Quality, and Service Quality. These dimensions jointly influence the success of OSS projects as they address different aspects that contribute to user satisfaction and system use (DeLone and McLean, 2003).

System Quality refers to the technical aspects of the software, including its reliability, efficiency, and performance. OSS projects with higher system quality are more likely to be used and appreciated by the community, resulting in higher adoption rates and user satisfaction (Crowston et al., 2003). Previous studies have measured System Quality through various means, including code quality metrics (e.g., bug density), and user evaluations (Blincoe et al., 2016). For our NLP artifacts, the most direct way of measuring the artifact quality is their benchmark performance, the recorded accuracy ran on specific benchmark datasets. This information is available for many model cards on Hugging Face Hub.

However, system quality is a multi-dimensional construct, and evaluation results alone may not be enough to capture the full picture. For NLP artifacts, the model structure, which further determines the hyperparameters, such as input dimension (max sequence length), can significantly contribute to the

overall system quality. NLP artifacts with better model structures are likely to perform more effectively in tasks. For example, the invention of the attention layer transformed state-of-the-art NLP and later models achieved significantly better benchmark performances (Vaswani et al., 2017). Another factor related to system quality is the size of the model. Larger models (with more layers and larger sizes) tend to require more computational resources to run, which can make them more unwieldy and inefficient.

Information Quality refers to the accuracy, completeness, and timeliness of the data and documentation provided with the OSS project. High information quality facilitates users to understand, use, and adapt the software to their needs, leading to increased adoption and satisfaction (Petter et al., 2008). In prior literature, Information Quality has been measured through both survey questionnaires and objective metrics such as the number of errors in the code (Borges et al., 2016). For our study, we can measure the information quality by extracting linguistic features from the model card description. An effective description would be informative and of sufficient length to cover the necessary details of technical and practical aspects. Additionally, it should include code snippets to demonstrate how to use the model, as previous literature pointed out that providing clear and helpful documentation information is extremely valuable for users and can therefore improve adoption (Steinmacher et al., 2015).

Service Quality focuses on the support and assistance provided by the OSS community, including user forums, troubleshooting, and collaborative development practices. A strong community can contribute to the success of an OSS project by fostering knowledge sharing, user engagement, and continuous improvement (Lakhani and Von Hippel, 2004; S. M. Lee and Lee, 2012). In the literature, Community Service Quality has been assessed through metrics such as the number of contributors, and the level of activity in user forums. For our NLP artifacts, we can extract the community service quality by its discussion post count and space count. Both features are native to the Hugging Face Hub, and they reflect the user-to-user support among the community well. For the field of NLP, the academic community is as important as the developer community, and we address this by crawling the citation count of the base model paper to reflect the community quality of academia

2.4. Control Variables

Prior research revealed and confirmed the role of geographical and cultural contexts, organizational backing, and industry contexts in the adoption of OSS. Recent and past studies examined developing countries and regional differences in OSS adoption (Gallego et al., 2008; Silva et al., 2023). Studies such as Poba-Nzaou et al. (2014), and Glynn et al. (2005) provide empirical evidence that these factors affect OSS adoption patterns. Therefore, overlooking these control variables could result in omitted variable bias, undermining the integrity of our conclusions. Informed by the consensus in the existing literature, we have selected three factors, geographical, organizational, and industry factors, as our control variables.

In the absence of user geographical data, we propose using the language of the NLP model as a proxy for capturing geographical and cultural effects. Given the constraints of our dataset, which lacks end-user organizational information, we will differentiate between models developed by organizations and those developed by individual contributors, as organizational backing improves credibility and facilitates adoption. For industry information, we will train a machine learning classifier to categorize NLP models based on their most probable industry applications, thereby capturing the effect of industry-related factors on adoption patterns.

2.5. NLP Task Typology

Natural Language Processing (NLP) has been through rapid evolution. A plethora of tasks exist to address the understanding, processing, and generating of human language. These tasks can vary significantly in their objectives, complexity, and techniques used. Despite numerous NLP artifacts being developed for different tasks, prior literature did not compare the artifacts on the task level, rather each study focused on an artifact alone. To better understand the impact of NLP tasks on the adoption and success of NLP artifacts, we posit adapting Ganagedara's task typology to our research model as a moderator.

Ganegedara (2018) classified NLP tasks into 2 categories: generation tasks and analysis tasks, as shown in the figure. Generation tasks involve creating new text based on specific input or context, such as machine translation or text summarization. On the other hand, analysis tasks involve extracting information, understanding, or processing existing text, such as sentiment analysis or named entity recognition. For analysis tasks, since the answer is usually known and

the complexity is relatively lower, evaluation results can be more meaningful and users could rely more on the system quality; for generation tasks, since their performances are harder to be measured by benchmark performance, and therefore users may rely more on service quality and information quality.

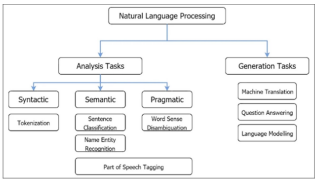

Figure 3. Ganegedara, 2018 NLP Task Taxonomy.

McGrath's Task Circumplex Model (1984) highlights the impact of task complexity on group behavior and performance in organizational settings. Although the model primarily focuses on task typology and group dynamics, it can provide valuable insights into the influence of task type on the behavior of the users of OSS NLP artifacts. As intellective tasks are different from divergent tasks, the latter are more ambiguous and riskier, and harder to assess the quality, it imposes increased cognitive challenges that users would face when deciding adoption. In our context of OSS NLP artifacts, analysis tasks are more intellective, while generation tasks are more divergent (creative).

Language translation tasks offer complex context that incorporates both analysis and generative task types depending on the length and complexity of translation tasks. While a single word or sentence translation may align more closely with analytical tasks, translating longer passages or engaging in speech-to-speech translation clearly leans towards generative tasks. Due to this reason, we employ a supplementary study to further explore the unique nuances of task typology within the language translation tasks, by focusing only on the adoption behavior among the translation models. By segmenting this specific task type, we aim to investigate the degree of analysis vs. generative tasks. We believe that this focused analysis will not only deepen our understanding of the overarching role of task typology in NLP tool adoption but also enrich the manuscript's core contributions by discussing the continuous effect of the task typology of analysis vs. generative.

Our research contributes to the literature by examining the impact of NLP task typology on the adoption and success of NLP artifacts in the context of OSS. By considering the dimensions of generation vs. analysis tasks, our study offers new insights into the preferences and behaviors of developers and users in the NLP domain. Since task-level comparisons have not been extensively explored in previous studies, our study makes an innovative contribution to the literature.

3. Research Model and Hypotheses Development

Our adaptation of the model includes the key constructs identified in the original model, such as system quality, information quality, and community quality. The three dimensions of quality subsequently determine the use and user satisfaction of the artifact. This relationship is further moderated by our novel dimension of task typology, which distinguishes between generation and analysis tasks in the NLP domain. Due to the difficulty of defining success as mentioned in the literature review, especially in the OSS context, the net benefits of these NLP artifacts are very hard to define, since OSS users may have different expectations of the outcomes. We limit our model to exclude the last stage of the original model of net benefits and focus solely on the adoption behaviors of use and user satisfaction. Net benefit could be measured in a more controlled environment. However, our research utilizes field data, and this level of control is not ideal. Our research model is shown in the figure.

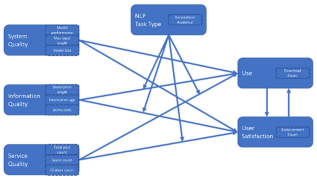

Figure 4. Proposed Research Model.

The system quality is captured using three variables: model performance, max input length, and model size. For the system quality construct, we hypothesize that:

H1: *System quality has a positive effect on use behavior and user satisfaction. Artifacts with better system quality will be used more and have better user satisfaction.*

The evaluation results depend on many factors, such as the dataset used, the performed task, and the evaluation method (for example, f-1 score, accuracy, loss, etc.). Many models report performance on more than one dataset, with multiple performance measures. For this reason, we use the best-recorded benchmark score from the model page. Since the scenarios for the intended use of those artifacts can vary a lot, there would always be a dataset where the model would perform poorly at. Due to this reason, the best-recorded score is more representative compared to the lowest or average score. We hypothesize accordingly:

H1a: *Model performance has a positive effect on download counts and endorsement counts. Artifacts with better evaluation scores will have more download*

counts and endorsement counts.

Max input length is captured by the model's hyperparameter which is included along with the model. The max sequence length is a good indicator of the quality and complexity of the artifact. Models with longer inputs can handle more contexts and are generally larger and more advanced. Although this benefit comes with a cost in computational resources, max input length is a good indicator of quality from a technical perspective for deep learning NLP models. We hypothesize accordingly:

H1b: *Max input length has a positive effect on download counts and endorsement counts. Artifacts with longer input lengths will have more download counts and endorsement counts.*

Model size refers to the size of the model file. As the model gets more complicated, it takes a larger file size to store all the extra layers and weights of the deep learning model. A larger model is generally more inconvenient to use, and developers develop distilled versions of the larger models that retain most of their performance while only taking a fraction of space. The model size is a good indicator of quality from a cost perspective for deep learning NLP models. We hypothesize accordingly:

H1c: *Model size has a negative effect on download counts and endorsement counts. Artifacts with smaller sizes will have more download counts and endorsement counts.*

The information quality is captured using three variables: description length, description age, and the presence of demo code. For the system quality construct, we hypothesize that:

H2: *Information quality has a positive effect on use behavior and user satisfaction. Artifacts with better information quality will be used more and have better user satisfaction.*

The length of the model card description is an important aspect of information quality. A good description needs to cover enough details to help users understand and use the artifact. A more comprehensive description typically provides greater detail about the model's functionality, purpose, and limitations. Since there is no spam or meaningless text on the model cards, and all model cards follow a certain archetype, length alone is good enough to capture the information density of the description. We hypothesize accordingly:

H2a: *The length of the model description has a positive effect on download counts and endorsement counts. Artifacts with longer descriptions will have more download counts and endorsement counts.*

The age of the description also represents a crucial aspect of information quality. As NLP is a rapidly evolving field, with models being frequently updated and uploaded, the timeliness of the description holds significant importance. Older descriptions may not contain the most up-to-date changes in how to use the model and may reflect that the model is deprecated or out of maintenance. Thus, description age can be an indicator of information quality for deep learning NLP models, with more recent descriptions being more likely to contain accurate and relevant information. We hypothesize accordingly:

H2b: *The age of the model description has a negative effect on download counts and endorsement counts. Artifacts with more recent descriptions will have more download counts and endorsement counts.*

The presence of demo code is another important factor that can significantly improve information quality for users by providing clear, hands-on examples of how to use the NLP artifact. Demo code helps users quickly understand the implementation and usage of the model, reducing the learning curve and increasing the likelihood of successful adoption. Therefore, the presence of a demo code is a good indicator of information quality for our study. We hypothesize accordingly:

H2c: *The presence of demo code has a positive effect on download counts and endorsement counts. Artifacts with demo code presence will have more download counts and endorsement counts.*

The service quality is captured using three variables: community post count, space count, and citation count. In the context of OSS, service is mainly provided by the community, and users seek service through the community; therefore, community quality is a core construct for our modified DeLone and McLean model. For the service quality construct, we hypothesize that:

H3: *Service quality has a positive effect on use behavior and user satisfaction. Artifacts with better Service quality will be used more and have better user satisfaction.*

The total post count serves as a measure of community activity. A higher number of posts indicates a more active community among its users. Models with more community activities mean posts asking for help may be more likely to be answered and more promptly. This reflects the service quality of OSS from a user-to-user perspective. We hypothesize accordingly:

H3a: *Community activity has a positive effect on download counts and endorsement counts. Artifacts with more community posts will have more download counts and endorsement counts.*

Space count represents the number of spaces in which the model is used. It is yet another indicator of community activity. Hugging Face Spaces is a

platform that hosts demos for deep learning artifacts. A higher space count signifies greater popularity among developers, reflecting the service quality of OSS from a popularity perspective. We hypothesize accordingly:

H3b: *The Hugging Face space count has a positive effect on download counts and endorsement counts. Artifacts that are used in more spaces will have more download counts and endorsement counts.*

Citation count is the third variable we consider. As many base NLP models, such as BERT for example, cited for 65k+ times (Devlin et al., 2018), have academic papers that are well-cited in the academic community across various disciplines, citation count is a good indicator of service quality from an academic perspective. For models with no paper related, we impute the missing citation count with 0. We hypothesize accordingly:

H3c: *The citation count in academia has a positive effect on download counts and endorsement counts. Artifacts with more citations in academia will have more download counts and endorsement counts.*

In our extension of the DeLone and McLean model, we consider the moderating effect of NLP task type. For the task type construct, we hypothesize that:

H4: *The NLP task type moderates the relationship between quality constructs and use/user satisfaction.*

For generation tasks such as question answering and summarization, since they are more subjective and complex by nature, measuring the actual performance of the model can be challenging; as a result, users likely would rely more on service quality and information quality to determine whether to use the artifact or not. For the analysis type of tasks, these quality dimensions play an even more critical role in shaping users' adoption decisions and satisfaction with the NLP artifacts. We hypothesize accordingly:

H4a: *The NLP task type moderates the relationship between system quality and use/user satisfaction. For the analysis type of artifacts, the relationship between system quality and use/user satisfaction is stronger than generation tasks.*

H4b: *The NLP task type moderates the relationship between information quality and use/user satisfaction. For the generation type of artifacts, the relationship between information quality and use/user satisfaction is stronger than analysis tasks.*

H4c: *The NLP task type moderates the relationship between service quality and use/user satisfaction. For the generation type of artifacts, the relationship between service quality and use/user satisfaction is stronger than analysis tasks.*

4. Data Collection

We will gather data from Hugging Face Hub (https://huggingface.co/models), a widely used OSS NLP platform where people upload, modify, and share models. We will extract the model card data from the website and obtain the variables for our research model. For our study, we focus on the 70,000+ models under the category of NLP tasks. After collecting the model cards, we will extract the variables: download/endorsement count, task type, evaluation score, max input length, model size, description length, age and demo code, total post count, and space count. We will train classifiers to identify evaluation scores and demo code presence.

5. Methodlogy

We will conduct a confirmatory factor analysis to confirm the internal consistency of our operationalization. After gathering the data, we use confirmatory factor analysis (CFA) to explore the relationship between the variables. This will ensure that our chosen variables are properly aligned with their respective constructs. We will report the factor loadings for each variable, which will provide insight into how well they relate to the underlying constructs. High factor loadings will indicate strong associations between variables and their respective constructs, thus validating our operationalization. Although the constructs of qualities are indeed correlated with each other, the loadings should show that within-group correlation is significantly higher compared to between groups. Next, we will run a linear regression model to estimate the relationship between the download counts and endorsement counts. Given the possibility of a long-tail distribution for download counts, we may need to apply a log transformation to address any potential skewness in the data. By analyzing the coefficients, p-values, and R-squared values to evaluate the strength and significance of the relationships between variables. To examine the robustness of our theoretical model, we will run multiple statistical models including and without the control variables. This dual approach will allow us to validate the strength and consistency of our initial findings. If a significant deviation in results exists, it may imply potential moderating effects could be investigated in subsequent studies.

6. Discussion

In our study, we expect to confirm our hypothesis that the three dimensions of quality contribute to the use and user satisfaction of OSS NLP artifacts. If confirmed, the implications of our findings will

have both theoretical and practical contributions to the understanding of adoption behavior in the OSS NLP context.From a theoretical perspective, our research extends the literature on OSS success by examining NLP artifacts and considering the task typology as a moderator for the use. The DeLone and McLean model has not been previously applied to NLP artifacts in OSS settings, and our study fills this gap by providing empirical evidence to support the model's generalizability. From a practical standpoint, our study reveals insights on the adoption behavior of OSS NLP artifacts, identifies the factors that influence user's decisions, and how the effects may vary based on task types. This insight can help developers and marketers tailor their strategies for different types of software. For example, within the context of NLP, generation tasks may require clear descriptions of the artifact's functionality and greater emphasis on community affirmation, compared to analysis tasks.

References

Alotaibi, R. S., & Alshahrani, S. M. (2022). An extended delone and mclean's model to determine the success factors of e-learning platform. *PeerJ Computer Science*, *8*, e876.

Barlas, P., Lanning, I., & Heavey, C. (2015). A survey of open source data science tools. *International Journal of Intelligent Computing and Cybernetics*, *8*(3), 232–261.

Blincoe, K., Sheoran, J., Goggins, S., Petakovic, E., & Damian, D. (2016). Understanding the popular users: Following, affiliation influence and leadership on github. *Information and Software Technology*, *70*, 30–39.

Borges, H., Hora, A., & Valente, M. T. (2016). Understanding the factors that impact the popularity of github repositories. *2016 IEEE international conference on software maintenance and evolution (ICSME)*, 334–344.

Chen, Y.-P., Chen, Y.-Y., Lin, J.-J., Huang, C.-H., Lai, F., et al. (2020). Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (alphabert): Development and performance evaluation. *JMIR medical informatics*, *8*(4), e17787.

Chowdhary, K., & Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603–649.

Crowston, K., Annabi, H., & Howison, J. (2003). Defining open source software project success.

Crowston, K., Howison, J., & Annabi, H. (2006). Information systems success in free and open source software development: Theory and measures. *Software Process: Improvement and Practice*, *11*(2), 123–148.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319–340.

DeLone, W. H., & McLean, E. R. (2003). The delone and mclean model of information systems success: A ten-year update. *Journal of management information systems*, *19*(4), 9–30.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64–73.

Fitzgerald, B., & Kenny, T. (2004). Developing an information systems infrastructure with open source software. *Ieee Software*, *21*(1), 50–55.

Gallego, M. D., Luna, P., & Bueno, S. (2008). Designing a forecasting analysis to understand the diffusion of open source software in the year 2010. *Technological Forecasting and Social Change*, *75*(5), 672–686.

Ganegedara, T. (2018). *Natural language processing with tensorflow: Teach language to machines using python's deep learning library*. Packt Publishing Ltd.

Ghapanchi, A. H., Aurum, A., & Low, G. (2011). A taxonomy for measuring the success of open source software projects. *First Monday*.

Glynn, E., Fitzgerald, B., & Exton, C. (2005). Commercial adoption of open source software: An empirical study. *2005 International Symposium on Empirical Software Engineering, 2005.*, 10–pp.

Gwebu, K. L., & Wang, J. (2011). Adoption of open source software: The role of social identification. *Decision support systems*, *51*(1), 220–229.

Hanč, J., Štrauch, P., Paňková, E., & Hančová, M. (2020). Teachers' perception of jupyter and r shiny as digital tools for open education and science. *arXiv preprint arXiv:2007.11262*.

Heron, M., Hanson, V. L., & Ricketts, I. (2013). Open source and accessibility: Advantages and limitations. *Journal of interaction Science*, *1*(1), 1–10.

Hollmann, V., Lee, H., Zo, H., & Ciganek, A. P. (2013). Examining success factors of open source software repositories: The case of osor. eu portal. *International Journal of Business Information Systems*, *14*(1), 1–20.

Howison, J., & Crowston, K. (2014). Collaboration through open superposition: A theory of the open source way. *Mis Quarterly*, *38*(1), 29–50.

Kross, S., Peng, R. D., Caffo, B. S., Gooding, I., & Leek, J. T. (2020). The democratization of data science education. *The American Statistician*, *74*(1), 1–7.

Lakhani, K. R., & Von Hippel, E. (2004). *How open source software works:"free" user-to-user assistance*. Springer.

Lee, S. M., & Lee, S.-H. (2012). Success factors of open-source enterprise information systems development. *Industrial Management & Data Systems*, *112*(7), 1065–1084.

Lee, S.-Y. T., Kim, H.-W., & Gupta, S. (2009). Measuring open source software success. *Omega*, *37*(2), 426–438.

Lenarduzzi, V., Taibi, D., Tosi, D., Lavazza, L., & Morasca, S. (2020). Open source software evaluation, selection, and adoption: A systematic literature review. *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 437–444.

Manaris, B. (1998). Natural language processing: A human-computer interaction perspective. In *Advances in computers* (pp. 1–66, Vol. 47). Elsevier.

McGrath, J. E. (1984). *Groups: Interaction and performance* (Vol. 14). Prentice-Hall Englewood Cliffs, NJ.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

Petter, S., DeLone, W., & McLean, E. (2008). Measuring information systems success: Models, dimensions, measures, and interrelationships. *European journal of information systems*, *17*, 236–263.

Poba-Nzaou, P., Raymond, L., & Fabi, B. (2014). Risk of adopting mission-critical oss applications: An interpretive case study. *International Journal of Operations & Production Management*, *34*(4), 477–512.

Ray, J., Johnny, O., Trovati, M., Sotiriadis, S., & Bessis, N. (2018). The rise of big data science: A survey of techniques, methods and approaches in the field of natural language processing and network theory. *Big Data and Cognitive Computing*, *2*(3), 22.

Rogers, E. M., Singhal, A., & Quinlan, M. M. (2014). Diffusion of innovations. In *An integrated approach to communication theory and research* (pp. 432–448). Routledge.

Rossi, B., Russo, B., & Succi, G. (2012). Adoption of free/libre open source software in public organizations: Factors of impact. *Information Technology & People*, *25*(2), 156–187.

Silva, D. G., Coutinho, C., & Costa, C. J. (2023). Factors influencing free and open-source software adoption in developing countries—an empirical study. *Journal of Open Innovation: Technology, Market, and Complexity*, *9*(1), 21–33.

Song, D., Vold, A., Madan, K., & Schilder, F. (2022). Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. *Information Systems*, *106*, 101718.

Steinmacher, I., Silva, M. A. G., Gerosa, M. A., & Redmiles, D. F. (2015). A systematic literature review on the barriers faced by newcomers to open source software projects. *Information and Software Technology*, *59*, 67–85.

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584–2589.

Subramaniam, C., Sen, R., & Nelson, M. L. (2009). Determinants of open source software project success: A longitudinal study. *Decision Support Systems*, *46*(2), 576–585.

Vasilescu, B., Yu, Y., Wang, H., Devanbu, P., & Filkov, V. (2015). Quality and productivity outcomes relating to continuous integration in github. *Proceedings of the 2015 10th joint meeting on foundations of software engineering*, 805–816.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38–45.