3-2021

How do users answer MATLAB questions on Q&A sites? A case study on stack overflow and MathWorks

Mahshid NAGHASHZADEH

Amir HAGSHENAS

Ashkan SAMI

David LO
Singapore Management University, davidlo@smu.edu.sg

Citation

NAGHASHZADEH, Mahshid; HAGSHENAS, Amir; SAMI, Ashkan; and LO, David. How do users answer MATLAB questions on Q&A sites? A case study on stack overflow and MathWorks. (2021). *28th IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER 2021)*.
Available at: https://ink.library.smu.edu.sg/sis_research/6863

How Do Users Answer MATLAB Questions on Q&A Sites? A Case Study on Stack Overflow and MathWorks

Mahshid Naghashzadeh
Department of Communications and Electronics Engineering
Shiraz University
Shiraz, Iran
mahshid.naghashzadeh@gmail.com

Amir Haghshenas
Department of Computer Science and Engineering and IT
Shiraz University
Shiraz, Iran
haghshenas.amir74@gmail.com

Ashkan Sami
Department of Computer Science and Engineering and IT
Shiraz University
Shiraz, Iran
sami@shirazu.ac.ir

David Lo
School of Information Systems
Singapore Management University
Singapore, Singapore
davidlo@smu.edu.sg

***Abstract*—MATLAB is an engineering programming language with various toolboxes that has a dedicated Question and Answer (Q&A) platform on the MathWorks website, which is similar to Stack Overflow (SO). Moreover, some MATLAB users ask their questions on SO. This paper aims to compare these two Q&A platforms to see what kind of questions are asked and how developers answer these questions in each platform. The result of our analysis on 80,382 MATLAB questions on SO and 266,367 questions on MathWorks show that MATLAB questions on topics ranging from the MATLAB software installation to questions related to programming received high votes and accepted answers on MathWorks. However, the questions about basics of programming such as plots, functions, and variables and questions on converting MATLAB code to other programming languages are very likely to receive answers on SO. Our detailed analysis on SO shows that users answer MATLAB questions with the same rate of the accepted answer as other popular programming languages like Java and Python, but the rate of unanswered questions and questions without an accepted answer for Simulink and the three most popular MATLAB toolboxes -- image processing, signal processing, and computer vision -- are very high. To analyze the evolution of MATLAB questions on SO, we studied 80,382 MATLAB questions using the SOTorrent dataset. The patterns in MATLAB questions' evolution are: 1) Most of the revisions to questions are text-related and not on code snippets. 2) Most of the code-related revisions were performed by the original poster (OP). 3) Non-original posters (Non-OPs) usually revise code snippets' appearance, while OPs usually revise code snippets' content and logic.**

Keywords—MATLAB, Stack Overflow, MathWorks, code snippets, evolution, question, revision

I. Introduction

Technical question and answer (Q&A) websites have changed how developers seek information on the web [1]. Stack Overflow (SO) is an ultimate platform that explicitly targets programmers and software engineers [2]. SO has emerged as one of the largest Q&A websites, where community members answer or participate in discussions to solve a problem [3].

MATLAB [1] is one of the most popular programming languages developed by MathWorks and Simulink is a MATLAB-based block diagram environment for multi-domain simulation. Widely used in all engineering and science fields, MATLAB is particularly famous for electrical engineering applications such as signal processing and control systems [4], and it is the foundation for many other tools [5]. Due to the significant number of MATLAB users globally and many MATLAB questions on Stack Overflow, and because there is already a platform dedicated to questions related to MATLAB called MathWorks, we decided to compare these two platforms and figure what kind of questions are asked in each platform and how other users answer these questions. We also performed a more detailed analysis on SO to determine how MATLAB questions are answered compared to other popular programming languages and how MATLAB questions evolve. Therefore, we started with 80,382 MATLAB questions from the SOTorrent dataset and narrowed them down to 32,161, which involved 70,450 revisions to see how users answer and revise MATLAB questions. In particular, we address the following research questions:

- **RQ1: What type of questions are more likely to get answered on MathWorks and SO?**

  On SO, MATLAB questions directly related to Simulink and three popular MATLAB toolboxes: image processing, signal processing, and computer vision are less likely to be answered or receive an accepted answer. We call questions related to Simulink and toolboxes "advanced questions." The questions about basics of programming: plots, functions, loop, variables, vectors, and matrices and questions on converting MATLAB code to other programming languages are very likely to receive answers, which we call basic programming questions. However, MATLAB questions on topics ranging from the MATLAB software installation to programming questions received high votes and accepted answers on MathWorks.

[1] Matrix laboratory

978-1-7281-9630-5/21/$31.00 ©2021 IEEE
DOI 10.1109/SANER50967.2021.00059

- **RQ2: Do MATLAB questions get answered on SO just as much as popular programming languages?**

  Users answer 86% of MATLAB questions, similar to answering other popular programming languages such as Python with 86% and Java, with 87% answered questions. The overall rate of accepting an answer is also similar to these programming languages.

- **RQ3: How do users participate in revising MATLAB questions in Stack Overflow?**

  Both OPs and non-OPs have almost the same participation level in terms of the number of revisions. 25.2% of OPs' revisions are code-related, and 74.8% of them are text-related. On the other hand, 12.5% of non-OPs' revisions are code-related, and 87.5% are text-related. We also find that 19% of all revisions on MATLAB questions are code-related, and 81% of them are text-related.

- **RQ4: How do users revise code snippets in MATLAB questions in terms of content?**

  Non-OPs usually revise code snippets' appearance (e.g., Move Text to Code Block), while OPs perform almost all of the content revisions on code snippets (e.g., Code Correction).

In short, the evolution of Stack Overflow questions follows some patterns. Most of the revisions on SO questions are text-related. Users usually revise questions to add code, correct the code, or put the source code of the question in the code block. OPs are more likely to revise the content of source codes, whereas non-OPs mostly revise code snippets' appearance. Based on our findings, a large percentage of questions on Simulink, image processing, signal processing, and computer vision are unanswered or do not have accepted answers. This indicates that the OPs of such questions are less likely to find solutions to their questions. Questions about basic MATLAB programming (e.g., questions about variables or loops) on SO are on-topic, but questions related to advanced topics on MATLAB are less likely to be answered. We want to tell users how they can improve the quality of MATLAB questions on SO and MathWorks. It can help shorten the answer receiving time and help users receive answers to their higher quality questions. The remainder of this paper is structured as follows. Section 2 identifies previous works that are related to our study. We explain data collection in Section 3. We report results and methodology in Sections 4, and Section 5 presents the threats to our study's validity. Finally, we highlight the conclusion in Section 6.

II. Previous Works

SO is a website that is collaboratively edited. Wang et al. [1] found that users performed more extensive than usual edits on the badge-awarding days. Users were more likely to perform text and small revisions if they performed many revisions in a single day. Jin and Servant [6] found that questions with more revises obtain more answers. Before receiving an answer, the most popular revises to questions were applied by the OP, on the body, and to clarify the meaning, and were not small. After receiving an answer, the most popular revises to questions were made by non-OPs, on tags, and adding related resources.

There are evidence-based guidelines that affect the success of questions. Adding code snippets to a question's body is strongly recommended [7]. The community answered review, conceptual and how-to questions more frequently than other kinds of questions. A possible reason is that code review questions can have more than one correct answer [8]. Also, Mi et al. investigated features of high-quality questions in SO and found out that the number of tags and code snippets are the most discriminative features [9].

III. Data Collection

We performed an empirical study over the SOTorrent dataset released on 23 August 2018 and 266,367 questions on MathWorks. SOTorrent is a dataset from the official SO data dump and the Google BigQuery GitHub (GH) dataset, providing the version history of Stack Overflow posts at the level of whole posts and individual text and code blocks [2]. We extracted questions containing MATLAB codes based on the tags containing "MATLAB," and 80,382 MATLAB questions were obtained. Our goal was to analyze and compare both code snippets and texts' evolution; thus, we consider code blocks and revisions for questions controlling factors. Therefore, we remove questions with no code blocks and get 56,516 questions. We also filter questions with no revision, leading us to 32,161 questions with at least one code block and one revision. We sorted all 266,367 questions on MathWorks in descending order based on their scores. Then, we filtered them based on whether they have an accepted answer or not.

IV. Results

In this section, we present the results of our research questions. First, we describe the approach used in analyzing data, and then we discuss the experimental findings.

RQ1: What type of questions are more likely to get answered on MathWorks and SO?

1) Approach

We applied a manual analysis on 96 selected MATLAB questions with the highest score and an accepted answer on MathWorks and SO. To calculate the size of selected questions, we used formula (1), where N is the population size (i.e., 46,040 MATLAB questions on SO and 266,367 questions on MathWorks), z is z-score, which is 1.96 for reaching 95% confidence level, p is population proportion (i.e., 0.5) and e is confidence interval (10%) [1]. Our first two authors manually studied and labeled the 96 questions on each Q&A website together and double-checked the labels. In the case of disagreement, we discussed the findings and explained the reason for our selection until we reached an agreement about the categories.

$$\frac{Nz^2p(1-p)}{e^2N+z^2p(1-p)} \quad (1)$$

2) Results

Based on our manual studies on the questions in Stack Overflow with the highest scores and an accepted answer, 64% of the questions are about MATLAB basic programming topics, including plots, functions, for loop, variables, vectors, and matrices. Also, 16% of questions are about converting MATLAB code to other programming languages, mostly Python, 10% about MATLAB performance and comparing it with other programs, and 10% of questions about other topics. We also investigated questions with the highest votes and an accepted answer on the MathWorks website, and it has been clear that 42% of these questions are around programming with MATLAB, for instance, plots, functions, and variables. Also, 14% of questions were about working with the MathWorks website's Q&A environment, 13% were about installing MATLAB software or additional toolboxes, 9% were about MATLAB software settings, 7% were around various versions of MATLAB's performance, 3% were concerning requests for adding features to MATLAB, and 10% of questions about other topics.

RQ2: Do Matlab questions get answered on SO just as much as popular programming languages?

1) Approach

We conducted a quantitative analysis to compare MATLAB questions with C, Python, and Java questions (as three popular programming languages) in terms of unanswered questions and questions without an accepted answer.

2) Results

In the quantitative analysis, we automatically analyzed the rate of unanswered questions and questions without an accepted answer between the four groups of MATLAB, C, Python, and Java (Fig. 1 and Fig. 2). Four groups have nearly the same rates of unanswered questions and questions without an accepted answer. Also, we wanted to understand which MATLAB questions are more likely to be answered on SO.

MATLAB has several toolboxes and a MATLAB-based block diagram environment for a multi-domain simulation called Simulink. We investigated the number of questions about Simulink and all MATLAB default toolboxes on SO. From all 11,780 questions about MATLAB default toolboxes and Simulink on SO, 10% (1,142) are computer vision, 11% (1,355) are signal processing, 20% (2,349) are Simulink, and 48% (5,672) are image processing questions. We extracted these numbers by automatically analyzing all tags containing the name of all MATLAB default toolboxes in questions.

We compared the rate of unanswered questions and questions without an accepted answer between four groups of MATLAB questions, i.e., Simulink and three MATLAB programming toolboxes, which are Computer vision toolbox, Image processing toolbox, and Signal processing toolbox (Fig. 3 and Fig. 4). Simulink questions have higher rates in unanswered questions and questions without an accepted answer than the other three groups, but in general, all four groups' rates are higher than overall MATLAB question rates.

The number of accepted answers in SO is a meaningful criterion that reflects the number of solved questions. Users should remind if they want to ask questions about basic programming problems in MATLAB, like plots, functions, and matrices, it is on-topic, but SO may not be an appropriate environment for advanced MATLAB topics such as Simulink, image processing, signal processing, and computer vision. Such questions may be on-topic for other Q&A websites like MathWorks.

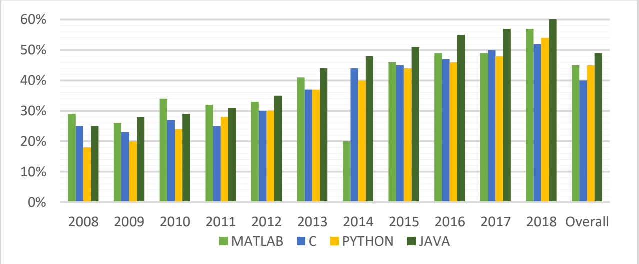

Fig. 1. The rate of questions without an accepted answer between MATLAB, C, Python, and Java

RQ3: How do users participate in revising MATLAB questions in Stack Overflow?

1) Approach

We categorized users into two groups: Original Poster of these questions (OPs) and other users (non-OPs). Also, we categorized revisions into two classes: code-related revisions and text-related revisions. Our primary motivation was to understand the participation level of each users groups in each class of revision.

2) Result

The result of the quantitative analysis on 32,161 MATLAB questions involving 70,450 revisions shows that OPs make 49.26% (34,700 revisions), and non-OPs make 50.74% (35,750 revisions) of the revisions on MATLAB questions. So, both OPs and non-OPs have almost the same participation level in terms of the number of revisions. We compared the number of code-related revisions and text-related revisions together, and we found that 19% (13,224 revisions) of all revisions on MATLAB questions are code-related, and 81% (57,226 revisions) of them are text-related.

In the next step, we analyzed revision classes in each group of OPs and non-OPs. 25.2% (8,750 revisions) of OPs' revisions are code-related and 74.8% (25,950 revisions) are text-related. In the same way, 12.5% (4,470 revisions) of non-OPs' revisions are code-related and 87.5% (31,280 revisions) are text-related. It means that OPs revise code blocks two times more than non-OPs. Fig. 5 shows the distribution of revisions performed by OPs and non-OPs in terms of class of revision, code-related, and text-related. In general, most of the revisions are text-related. Although non-OPs have acceptable participation in revising posts, OPs still have more enormous contributions in code-related revisions. Another significant result is that 30% of questions do not include any code blocks, whereas, based on SO

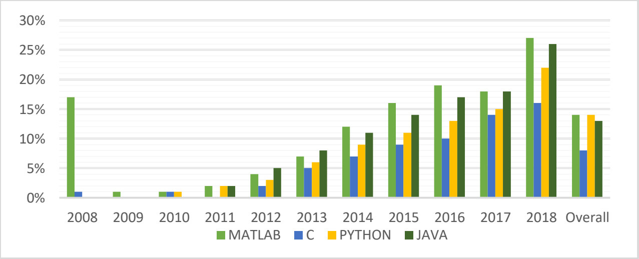

Fig. 2. The rate of unanswered questions between MATLAB, C, Python, and Java

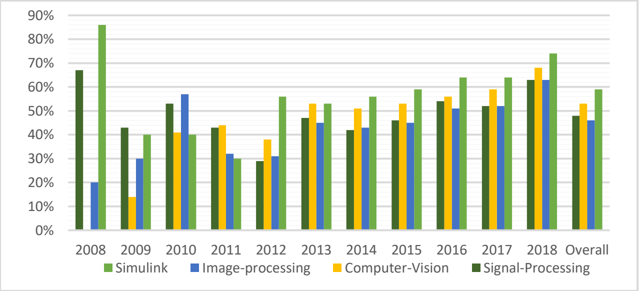

Fig. 3. Bars show the questions' rate without an accepted answer for each of the four groups of MATLAB advanced questions.

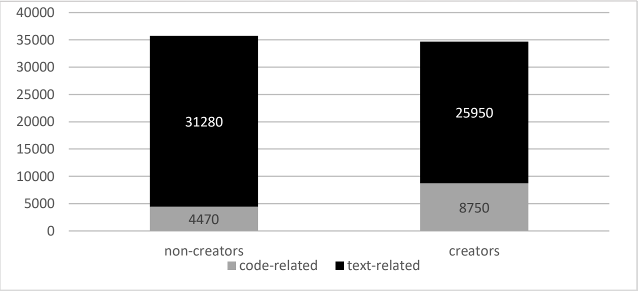

Fig. 5. Distribution of revisions made by OPs and non-OPs in terms of class of revision, code-related and text-related.

guidelines, users should place codes into code blocks of questions.

RQ4: How do users revise code snippets in MATLAB questions in terms of content?

1) Approach

We conducted a qualitative analysis to find out for what reasons users revise MATLAB questions on SO. Our primary motivation was to understand if there is any pattern in revising code snippets. We randomly sampled 384 questions of our dataset and accurately studied all of the questions' revisions. To calculate the size of the random sample, we used formula (1), where N is 32161, z is 1.96 for reaching 95% confidence level, p is population proportion (i.e., 0.5), and e is confidence interval (5%) [1]. In RQ1, we used a 10% confidence interval to understand interesting topics for MathWorks and SO users. The lower confidence interval would lead to a larger sample size, and questions without high scores would place in the sample. But, in this RQ, to highlight the reasons behind MATLAB questions' revisions, we had to use a larger sample to find those reasons.

We studied all of the 384 questions with all of their revisions to understand why users revised code snippets of questions. Then we examined all of these reasons, and we found that users revise questions for a limited number of reasons. So, inspired by the categories of significant reasons behind revisions reported in [1], we came up with the categories presented in Table 1. For instance, whether users added a new code block is categorized as 'Code Addition' or changed a variable's value as 'Code Correction.'

1) Results

Fig. 6 shows the distribution of significant reasons behind revisions. The first three most common reasons behind revisions are adding new code to a question (Code Addition), Code Correction, and Move Text to Code block with 32%, 18%, and 16% of frequencies, respectively. (Move Text to Code block means the user inserts the question's code into the text block, and after the question is edited, the code is put in its correct place, i.e., the code block.)

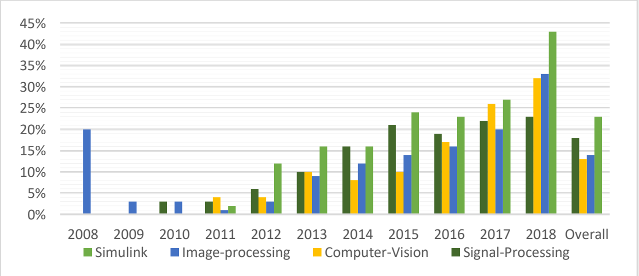

Fig. 4. The rate of unanswered questions for each of the four groups of MATLAB advanced questions.

We compared reasons that cause code snippets to be edited in MATLAB questions between OPs and non-OPs (Fig. 7). Non-OPs are more likely to help with editing the appearance of code snippets since 87.5% of all non-OPs revisions include Code Style Improvement and Move Text to Code block (i.e., appearance editing), and 12.5% are applied to the content of code snippets. Whereas almost all code snippets content revisions are performed by OPs with a rate of 92.1%, and only 7.9% of them are applied to code snippets' appearance. OPs fix a bug in a code block (Code Correction), change functions or logic (Code Functionality/Logic Improvement), change the name of variables, add or remove the comment in code (Code Readability improvement), add or remove code in questions (Code Addition/Removal).

Non-OPs have less contribution in revising the content of code snippets. One possible reason is that editing content of code snippets in a question needs in-depth knowledge about that code, and usually, OPs themselves have such knowledge. For instance, add variables or changing their values need prior knowledge about questions, and these types of revisions depend on what OP means to achieve from questions. For instance, for other types of revisions, fix a bug in code snippets (code correction), or improve code readability, non-OPs have enough information. We believe that if non-OPs contribute to making such revisions, they can speed up the evolution of questions, they can improve the quality of SO content, and OPs will receive answers for questions in a shorter time.

TABLE I. MAJOR REASONS THAT CAUSE REVISIONS

	Revision Reason	Definition
1	Code Correction	Addition of variables or change the value of variables, fix a bug in a code block
2	Code Functionality/Logic Improvement	Changing functionality or logic in code blocks
3	Code Readability Improvement	Changing the name of variables, adding or removing comments in code blocks
4	Code Style Improvement	Adding or removing space/lines
5	Code Addition	Adding lines of code or new code blocks to a question
6	Code Removal	Removing lines or entire code block
7	Move Text to Code Block	Placing source code of questions from text block into the code block

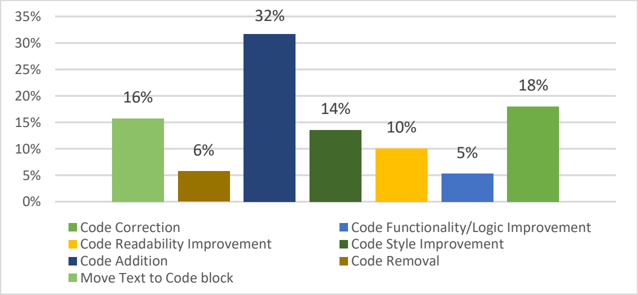

Fig. 6. The distribution of the reasons behind a revision based on randomly sampled data

V. Threats to Validity

A. External validity

Threats to external validity relate to the generalizability of our findings. In this research, we studied MATLAB questions in Stack Overflow and MathWorks; thus, our findings may not be generalizable to other Q&A websites and other programming languages.

We performed a qualitative analysis in RQ1 to find which MATLAB questions answered better on SO and MathWorks. We sampled 46,040 questions on SO and 266,367 questions on MathWorks with a 95% confidence level and a 10% confidence interval. We also conducted the same qualitative study in RQ4 to find the main reasons behind revisions of MATLAB questions in SO because it was impossible to study 32,161 questions manually with all of their revisions. Therefore, we randomly sampled the questions with a 95% confidence level and a 5% confidence interval (we ended up studying 384 randomly sampled questions and all of their revisions in RQ2).

B. Internal validity

Threats to internal validity relate to the experimenters' bias and error in qualitative analysis in RQ1 and RQ4. Questions were labeled and double-checked by the first two authors and validated by other authors to reduce bias and error in qualitative analysis in RQ1. To reduce bias and error in RQ4, we repeated qualitative analysis two times, and the results are mean of these statistics, i.e., in general, we studied 768 MATLAB questions with all of their revisions.

This study recognizes questions based on tags containing related words, but there may be MATLAB questions without including tags that we investigated on SO or MathWorks. This may cause a threat to the validity of our study.

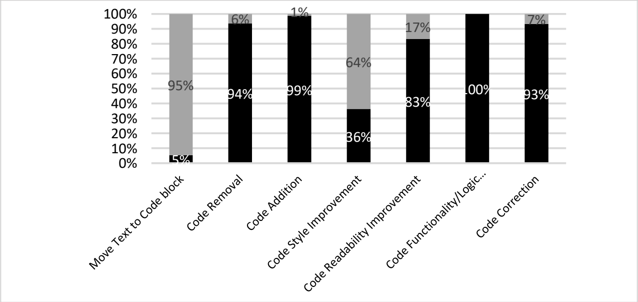

Fig. 7. Compare revisions made by Ops (black bars) with revisions made by non-Ops (gray bars).

VI. Conclusion

This paper analyzed 80,382 MATLAB questions on Stack Overflow and 266,367 questions on MathWorks to determine how users answer MATLAB questions on these two Q&A platforms. We chose MATLAB because it is a widely used tool in different fields, and there are many MATLAB questions asked on Stack Overflow. Also, because MathWorks is already a platform dedicated to MATLAB, we decided to investigate how Stack Overflow performs as a secondary Q&A platform for MATLAB questions.

In the first research question, we find that MATLAB basic programming questions are more likely to receive an accepted answer and a high score. However, questions on topics ranging from the MATLAB software installation to programming questions received high votes and accepted answers on MathWorks. In the second research question, we find that the rate of unanswered questions and questions without an accepted answer in MATLAB questions is nearly the same as C, Python, and Java. Advanced MATLAB topics' questions such as Simulink, image processing, signal processing, and computer vision are less likely to be answered on Stack Overflow. The third research question shows that 19% of revisions are code-related, and 81% are text-related. OPs revise code blocks two times more than non-OPs. Also, 30% of questions do not include any code blocks. To find significant reasons behind revisions in the fourth research question, we studied 384 randomly sampled questions with all of their revisions. The three most common reasons behind code-related revisions are Code Addition, Code Correction, and Move Text to Code Block. Non-OPs usually revise the appearance of code snippets, while OPs usually revise the content of code snippets.

References

[1] S. Wang, T. H. P. Chen, and A. E. Hassan, "How Do Users Revise Answers on Technical Q&A Websites? A Case Study on Stack Overflow," IEEE Transactions on Software Engineering, pp. 1–19, 2018.

[2] S. Baltes, C. Treude, and S. Diehl, "SOTorrent: Studying the origin, evolution, and usage of stack overflow code snippets," IEEE/ACM 16th International Conference on Mining Software Repositories (MSR), Montreal, QC, Canada, pp. 191–194, 2019.

[3] M. Asaduzzaman, A. S. Mashiyat, C. K. Roy, and K. A. Schneider, "Answering questions about unanswered questions of Stack Overflow," 10th Working Conference on Mining Software Repositories (MSR), San Francisco, CA, USA, pp. 97-100, 2013.

[4] H. Moore, MATLAB for Engineers, NJ, Upper Saddle River: Pearson/Prentice-Hall, 2007.

[5] W. J. Palm, Introduction to MATLAB 7 for Engineers. New York: McGraw Hill Professional, 2005.

[6] X. Jin and F. Servant, "What Edits are Done on the Highly Answered Questions in Stack Overflow? An Empirical Study," IEEE/ACM 16th International Conference on Mining Software Repositories (MSR), Montreal, QC, Canada, pp. 225-229, 2019.

[7] F. Calefato, F. Lanubile, and N. Novielli, "How to ask for technical help? Evidence-based guidelines for writing questions on Stack Overflow," Information and Software Technology, vol. 94, pp. 186–207, 2017.

[8] C. Treude, O. Barzilay, and M. A. Storey, "How do programmers ask and answer questions on the web? (NIER track)," In Proceedings of ICSE 2011, New York, NY, USA, pp. 804-807, 2011.

[9] Q. Mi, Y. Gao, J. Keung, Y. Xiao, and S. Mensah, "Identifying Textual Features of High-Quality Questions: An Empirical Study on Stack Overflow," 24th Asia-Pacific Software Engineering Conference (APSEC), pp. 636–641, 2017.