# Extracting Insights about Deep Learning Posts from Stack Overflow

Ritik Arora
University of Waterloo
ritik.arora@uwaterloo.ca

Varshanth Rao
University of Waterloo
varshanth.rao@uwaterloo.ca

Igor Henrique Nicacio Braga
University of Waterloo
ihnicaci@uwaterloo.ca

## ABSTRACT

In recent years, the field of deep learning has garnered more interest mainly due to the availability of vast amount of data and exponential improvement in processing power of computing devices. To share and expand their competency in a field like deep learning, Question and Answer sites provide essential support for developers and researchers. In this paper we investigate and extract data from a popular Q & A site, Stack Overflow to gather insights about the questions being asked and the corresponding responses on the topic of deep learning based on a set of tags. We gathered data related to posts by querying the Stack Exchange Data Explorer regarding information about the users and votes for each posts. We examined them to understand about how the designation and the reputation of user affect the quality of answers and how much time does it usually take to have an acceptable answer. Through our research, we attempted to answer 3 important research questions which we address in detail related to the relationship between the designation of the users and the reputation of the users on the quality of the question and answer posts. Furthermore, we derive insights regarding how quick one should expect good quality answers to arrive when a deep learning question has been posted.

## 1. INTRODUCTION

With the advancement of technology particularly in processing power offered by GPUs and the vast amount of data available, Deep Learning has started to progressively gain traction in recent years. Because of this, Deep learning has attracted interest not only from researchers, but also from software industry. Due to this high demand for deep learning, a lot of jobs are being created in this domain as well as a lot of scope for new research has been generated. But for the development of skill in any new technology, support for issues and questions in that area is essential. To get this support and gain more expertise in a domain, users tend to rely on Question and Answer sites like Stack Overflow where they can post their issues or help others in resolving their issues. Stack Overflow community is responsible for adhering to the quality standards of the questions and answers which users can do by upvoting or downvoting any question post along with its answer.

Many challenges exist to help provide proper support for the new users in a particular domain. New, as well as old users could be interested in learning about trending topics in deep learning. How much credibility does a user who is asking or answering a question has? How does a user's overall reputation relate to the type of questions or answers he or she is posting about topics in deep learning domain? Deep learning has garnered interest and we are witnessing its penetration into the industry as well. What is the distribution of user profession in the deep learning community and does their profession affect the quality of the posts that they upload? Given that question posts are tagged, could we predict the effective response time for answers?

To derive these and some more insights about Deep learning, we mined Stack Overflow for the data about posts, users and votes for each posts. Stack Overflow provides a tool - Stack Exchange Data Explorer, which helps in executing SQL queries to extract data from Stack Overflow database. We collected a list of 30 tags which were most related and specific to the deep learning domain. Additionally, we parsed the AboutMe field of a user to map profession related keywords to a list of 8 broad designations categories. We mined the data to extract question posts based on the tags and their corresponding answers. We obtained 63,177 questions and 29,602 answers posts. The research questions which we attempt to answer are described below.

- RQ1.1: How does the designation (e.g. student, researcher, developer, professor, etc.) of the user affect the quality of questions and answers posted?
- RQ1.2: How does the reputation (beginner or expert) of a user affect the quality of answers (no of up-votes)?
- RQ2: How long does a questioner have to wait before he/she can expect acceptable answers for his/her deep learning tagged questions?

Based on the results we observed that across all tags, Industry professionals followed by academic students are the most active users in deep learning domain. Surprisingly, the user community related to the designation, "Academia-Pedagogical", have very less involvement in the topics which are more theoretical. We were able to distinguish between tags on the basis of the confusion level and presence of community support. For example, the tags such as caffe and convolution have more clarity within the community as compared to the tags like tensorflow-gpu, lstm, rnn which are less explored and can be viewed as arcane. Another surprising discovery was that most of the questions and answers which were posted, lie in the low score range and are being asked and answered by users with reputation in the low range. Questions and answers belonging to high score range were also being asked and answered by low to medium reputation users. We derived an additional parameter, the Smoothed Weighted Upvote-Downvote Ratio (SWUDR), to observe more granularity amongst the score ranges for both questions and answers. Lastly, we found that for most of the tags, the 3rd most upvoted answer takes the least amount of time to be answered. More insights and inferences are detailed in the relevant results section of the paper.

The rest of the paper is outlined as Section 2, which describes the related work being conducted on similar lines. We focused on research papers that extract various insights with similar approaches. Section 3 describes about the methodology we implemented to mine the data, pre-process and represent it. In Section 4 we present our results along with some of the insights we were able to derive from these outcomes, followed by threats to validity presented in Section 5. Finally, we conclude our research in Section 6 and outline some of the areas where further research could be done.

# 2. RELATED WORK

## 2.1 Stack Overflow

Stack Overflow is a collaborative learning environment where users can ask and answer questions related to software, programming languages, concepts, platforms, operating systems and etc. It receives more than six thousand questions on a daily basis [8]. To answer a question or post one, the user must create an account if she does not already have one. Users must input their information such as name, designation, date of birth, and other personal information. Even though Stack Overflow demands several information to create an account, those fields aren't publicly due to privacy concerns. We discuss methods to circumvent the lack of information to some extent, in the following section.

Once with an account, a user can post, answer, up-vote, down-vote any question. They can also up-vote and down-vote answers (quantity of up-votes minus quantity of down-votes constitutes the score of a question/answer), however, only the user who posted can mark the question as answered. The later will contain up to five key words, known as tags, which could be specific to a programming language (C++, Python, Java), a framework (AngularJS, ReactJS) or even broader areas such as Deep Learning which is the focus of this paper [11].

## 2.2 Background

There have been several papers that also tried to extract useful information from Stack Overflow. For example, in [6] researchers try to investigate why women are often impeded from contributing to questions and answers. So they studied how the presence and how often women contribute to these posts. They found out that women are more likely to contribute to posts made from other women. Similarly, [9] also engage in research related to gender; however, instead of investigating solely women, they try to understand the gender disparity in Stack Overflow by using gender guessers. They observed that combining different data sources yielded the best results. While [9] does not mention the source of their data, [6] used Stack Overflow Data Exchange (SEDE), the same used by this paper.

Furthermore, in [11] the authors investigate the relationship between programming language and its experts to identify what languages should be recommended to newcomers (e.g. students) and to experts. We used a similar approach as [11] when it comes to writing scripts to extract further information from the data. [5] focus on integrating Stack Overflow to IDEs, more specifically Eclipse, where instead of having to access Stack Overflow data in a browser, the developer is able to access it through the IDE itself.

[10] asks the question "is programming knowledge related to age?" They study the relation of age and effect on programming knowledge and how age relates to a developer's skill set. They found out that there is not a strong correlation between age and scores in specific knowledge areas. Additionally, they observed that into the age of 50, programmer's reputation score tended to increase while developer in their 30's tend to focus on fewer areas relative to those who were younger or older. Unlike the previous papers, the authors extracted their data from the 2013 MSR Mining challenge PostgreSQL data dump [4]. Further on our research, we try to relate our results to the ones found by [10].

Most of these papers target specific questions about programmers, programming languages, topics, and habits. Interestingly, [12] focus on why there are so many unanswered questions on software information sites. After using data mining techniques their results showed that lack of interest of the user community played a big role on those unanswered questions.

Lastly, [2] proposed a machine learning regression model for predicting the score (quality) of a question on Stack Overflow. After examining several features they used 16 factors related to question format (ratio of body length to paragraphs), content polarity and subjectivity (positive, neutral or negative). They found out strong correlations between number of tags, question's length, accepted answer score and the scores of the questions.

It is noteworthy that researchers have extensively used Stack Overflow to gain insights from developers community [10], to study which questions are more relevant and why [11], to question why we do not see as many women contributing to Stack Overflow ([6], [9]), to investigate how age and developer's skills are related [10], to build recommendation systems to help newcomers and experts ([2], [13]) and to increase awareness of potential license violations [3]. We also contribute to the same by answering several research questions previously mentioned.

# 3. METHODOLOGY

As mentioned in the introduction section, we used the Stack Exchange Data Explorer (SEDE) to mine the Stack Overflow data. SEDE is a querying platform from which users can execute SQL like queries to get information about the posts and users on the supported online forums. To address our research questions it was essential for us to filter exactly the data required.

The key information we wanted collectively was related to the users, the posts and the votes table. More specifically, for the research questions we required information regarding the designation, reputation and post details of the questioners or answerers of all users in the deep learning community. We first had to isolate and define the constituents of the deep learning community at Stack Overflow. The posts on

| Tag List |
|---|
| attention-model, autoencoder, backpropagation, batch-normalization, bias-neuron, caffe, conv-neural-network, convolution, deconvolution, deep-dream, deep-learning, deep-residual-networks, deeplearning4j, feed-forward, keras, keras-layer, lstm, max-pooling, neural-network, pytorch, recurrent-neural-network, resnet, rnn, sequence-to-sequence, tensor, tensorboard, tensorflow, tensorflow-gpu, tensorflow-slim, torch |

Table 1: Tag List - 30 Unique Tags Used

Stack Overflow are topic labeled using "Tags" where each question can comprise of a maximum of 5 tags. We found the basic tag of "deep-learning" and queried to find the top 10 users contributing to this tag. We performed a breadth first search on all the tags that these users posted their questions and answers in and selected the tags exclusively related to deep learning. Since some terminologies of deep learning also seeps into other communities of machine learning (such as gradient descent, epochs etc.) we excluded these tags to narrow down the scope of our research. We finalized a set of 30 tags which is the primary basis for our entire experiment, which is listed in Table 1.

For RQ1.1, we queried the SEDE for all posts of questions and answers filtered by the shortlisted tags. Unfortunately, the designation of the author of the post was not accessible even though Stack Overflow requires this while signing up to the community to ask and answer questions. The Users table had an AboutMe field where we found many users had posted their profile. We opined that mining the AboutMe field would yield the designation of at least 30% of the community. We classified the different designations into 7 categories according to a designation-keywords map which is shown in Table 2. The mappings were processed in order (top to bottom) to avoid overlapping designation assignments. To eliminate a possible threat to validity, we filtered out the posts from users who did not have a UserId associated with them which indicated that their user profile has expired or is no longer valid, hence rendering their information obsolete. Since a question post can be associated with many answers, we considered only the most upvoted answer for RQ1.1. Ultimately, we obtained 63,177 question posts and 29,602 answer posts. We present the information by grouping the output by tag and within each tag grouping by designation.

For RQ1.2, we used the same queried output from the SEDE as RQ1.1. We focus RQ1.2 on the distribution of the users across the questions and answers. For addressing the answers portion of RQ1.2, we take the top 3 most upvoted answers for each question to get a more holistic view of the distribution of answerers amongst the different score ranges of the posts. For both the question and answer posts, we first group by tag. The distribution of the questions and answers within each tag are then categorized into 3 score ranges which is derived from the "Score" field of the posts. The score ranges for each tag are derived by splitting the difference between the maximum and minimum score of the questions and answers within the specific tag into 3 difference chunks. Similarly within each tag, we define a user community of questioners and answerers and group them into reputation ranges. The reputation ranges for each tag are derived by splitting the difference between the maximum and minimum "Reputation" of the users into 3 difference chunks. For each score range of the questions and answers, the distribution of the questioners and answerers respectively local to each tag were distributed amongst the reputation ranges corresponding to their reputation values. For reasons which we will explain in the RQ1.2 section, we also calculate an additional metric analogous to "Reputation" which we call as the Smoothed Weighted Upvote-Downvote Ratio (SWUDR)

The SWUDR is calculated as below:

Let
U represent User,
PT represent PostType (Question or Answer),
UP represent Upvotes within Deep Learning Posts,
DN represent Downvotes within Deep Learning Posts,
PP represent Participation

$$\boxed{SWUDR(U,PT) = \frac{(TotalUP(U,PT)+1) \times PP(U,PT)}{((TotalDN(U,PT)+1) \times MaxPP(PT))}}$$

We represent the same data as described above but instead of "Reputation" as the user metric, we use the "SWUDR" as the primary user metric and use SWUDR ranges instead of reputation ranges.

We use the top 3 most upvoted answers data which we used as input for the representation for RQ1.2, for RQ2 we obtained 51,413 answer posts using the above data. We pre-process the data by calculating the response time as the time difference between answer creation date and the question creation date. The data is then flattened so that each row represents a question, the top 3 answers and the corresponding response times. For RQ2, this information is presented by grouping the question-response tuples according to the tag and then processing the mean response times for the first, second and third most upvoted answers.

We used SEDE to query the Stack Overflow Data Set, we used the Pandas library to pre-process the queried data and Tableau Software to visualize our results.

## 4. RESULTS

To discuss and visualize the results we constructed diagrams for each of the research questions. This section explains in detail the insights obtained from our queries/scripts and diagrams for each research question. We start with our first research question "How does the effect of previous experiences/reputation reflect on the quality of questions and answers?" which is divided into two sub-questions.

### 4.1 Research Question 1.1

*How does the designation (e.g. student, researcher, developer, professor, etc.) of the user affect the quality of questions and answers posted?*

Our first research question (RQ1.1, see above) uses designation strength as a metric. The Designation Strength denotes

| Designation (Processed In Order) | Keywords Used |
|---|---|
| Academia - Professional | research engineer, data scientist, research scientist |
| Professional | developer, engineer, designer, analyst, contractor, software engineer, software developer, data engineer, graphic designer, cs engineer, computer engineer, consultant, architect, manager, team lead, development, professional, founder, leader, tech lead, technical leader, work on, dev, freelance, work at, company, specialist, startup |
| Academia - Ambiguous | researcher, computer scientist, mathematician, bioinformatician, statistician, physicist, neuroscientist, biologist, scientist |
| Academia - Pedagogical | professor, lecturer, teacher, teach |
| Academia - Student | student, freshman,sophomore, grad, b.e , mtech, bachelor, phd, ph.d, school, studying, thesis, btech academic, college, teaching assistant, university, data science, computer science, intern, msc, postdoc, engineering, research, masters |
| Professional - Ambiguous | programmer, coder |
| Online-User-Profile-Present | http, www |
| Unknown | * |

Table 2: Designations - Keywords Mapping

the number of question or answer posts that were asked or answered by those users who were categorized into that designation. Figure 1 shows the amount/percentage of each designation from the entire data. As we can see, the designation "Unknown" dominates the population. We hypothesize that these users either create accounts just to get a few questions answered, do not wish to enter designation information in the AboutMe section as it may be redundant or just want to stay anonymous. After manual inspection we found that the AboutMe section of most "Unknown" designations were left blank or contained information irrelevant to any form of profession. For Figure 2 we ignored the Unknowns designation for better visualization. This is further discussed in the threats to validity section.

Based on RQ1.1 we were able to observe that industry professionals are the most active in all the fields of deep learning, ranging across all tags as well as questions and answers. (See Figure 2) We hypothesize that this is the case because when such topics become part of a developers daily routine/job, she will have more opportunities to come across doubts, questions and solutions related to those topics. In comparison to academia-students who might be involved in multiple projects and classes at the same time, which could involve multiple topics.

Furthermore, the second group most active in both questions and answers were academic students category. As mentioned earlier, even though students are involved in multiple projects and classes, they normally are in the process of learning new topics; hence, they ask many questions. On the other hand, many students have strengths (topics that they feel comfortable with). That could explain why academia-students are the second most active group in topics such as TensorFlow and Neural Network.

Surprisingly, in theoretical topics such as neural networks, conv-neural-network, back propagation, and lstm, the involvement of users of the Academia-Pedagogical designation was minimal in the answer posts. We infer that they are either not on Stack Overflow or they are not actively participating in such discussions. Moreover, we can deduce from even distribution of questions and answers in tags such as tensorflow, neural-network, deep-learning, keras, and torch, that there exists a strong community behind these topics.

Tags such as caffe, convolution, backpropagation, and conv-neural-network have more answers than questions. Therefore, we hypothesize that the deep learning community are more proficient or have a conceptually clear understanding in these topics. This graph only shows the answers that was most upvoted for each question and not the total amount of answers, which could be different. However, we believe that this difference should not affect our hypothesis due to the results found from our second research question (RQ2).

On the other hand, tags such as rnn, recurrent-neural-network, tensorflow-gpu, lstm (Long Short Term Memory) have more questions than answers; thus, we hypothesize that for these topics, the user community is less certain, has more doubt or more confusion. Additionally, we could say that users have interest in these topics; although, there aren't sufficient community support available. This is surprising because there are good evidence that Deep Learning has grown in both academia and industry [7]. For instance, from Figure 2 we can observe that Academic-Professionals and Academia-Ambiguous people are equally involved in deep learning. Even though they are not majority, we can still observe a considerable number of researchers which are interested in the topic, which explains, at least in part, the increase in research in this area.

It is not surprising that students in academia are asking more questions than answering them. The latter are mostly answered by professionals. As we mentioned earlier, professionals might feel more comfortable than any other

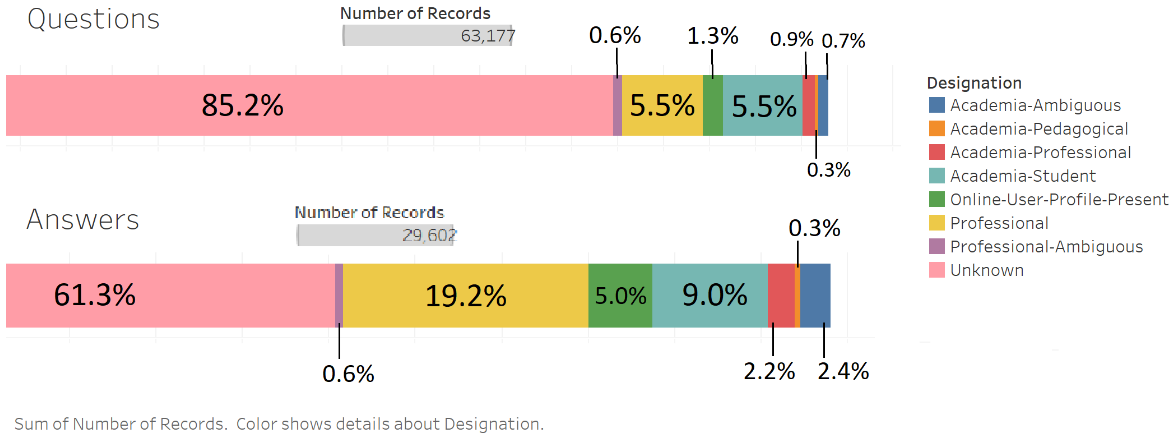


Figure 1: Question and answer posts distribution the different designations

designation because they work with these topics on a daily basis to deploy applications to the outside world. Similarly, Academia-Students post more questions compared to other designation in topics such as, conv-neural-network, backpropagation, lstm, and pytorch. We can conclude that these topics have not fully transitioned to industry, or if they did, professionals aren't as confident with them.

Lastly, we found that PyTorch was launched in late 2016 and it is a topic of interest among academic students and not as much among professionals. We believe that since PyTorch is a new library, it is not used as much in industry; or if it is used, it is in small numbers. It could also be the case that TensorFlow has proliferated as the industry standard with a large community base and hence most industries that make use of Deep Learning algorithms use TensorFlow. However it is not clear if companies, institutions and students will adopt which library or package and we do not discuss this further because it is out of the scope of this research.

## 4.2 Research Question 1.2

*How does the reputation of a user affect the quality of answers?*

As discussed in the methodology section, we categorized both question and answer by tags and then calculated the score range for questions which were further divided into range of reputation of users who have either asked or answered those posts. In Figure 3a and Figure 3b, we observed a general trend for all the tags. Most of the questions and answers fall in the lower score range and have been asked and answered respectively by users that belong to a low reputation range. But a surprising discovery was that questions and answers that belong to the high score range are even being asked and answered respectively by people with low to medium reputation range.

Another interesting fact is that those users whose reputation lies in the high reputation range have asked most questions that belong to a low score range. This was a bit confusing since ideally we would expect users with high reputations to be asking high scored questions. We hypothesized that the reputation of users that we mined from Stack Overflow database is an accumulated representation that users have earned by asking and answers posts across multiple domains and not just deep learning. To inspect in more depth, we calculated a factor local to deep learning tags called Smoothed Weighted Upvote-Downvote Ratio (SWUDR), which was explained in the methodology section, and conducted the same experiment as descried above. We can see the results with SWUDR factor in Figure 3c and Figure 3d.

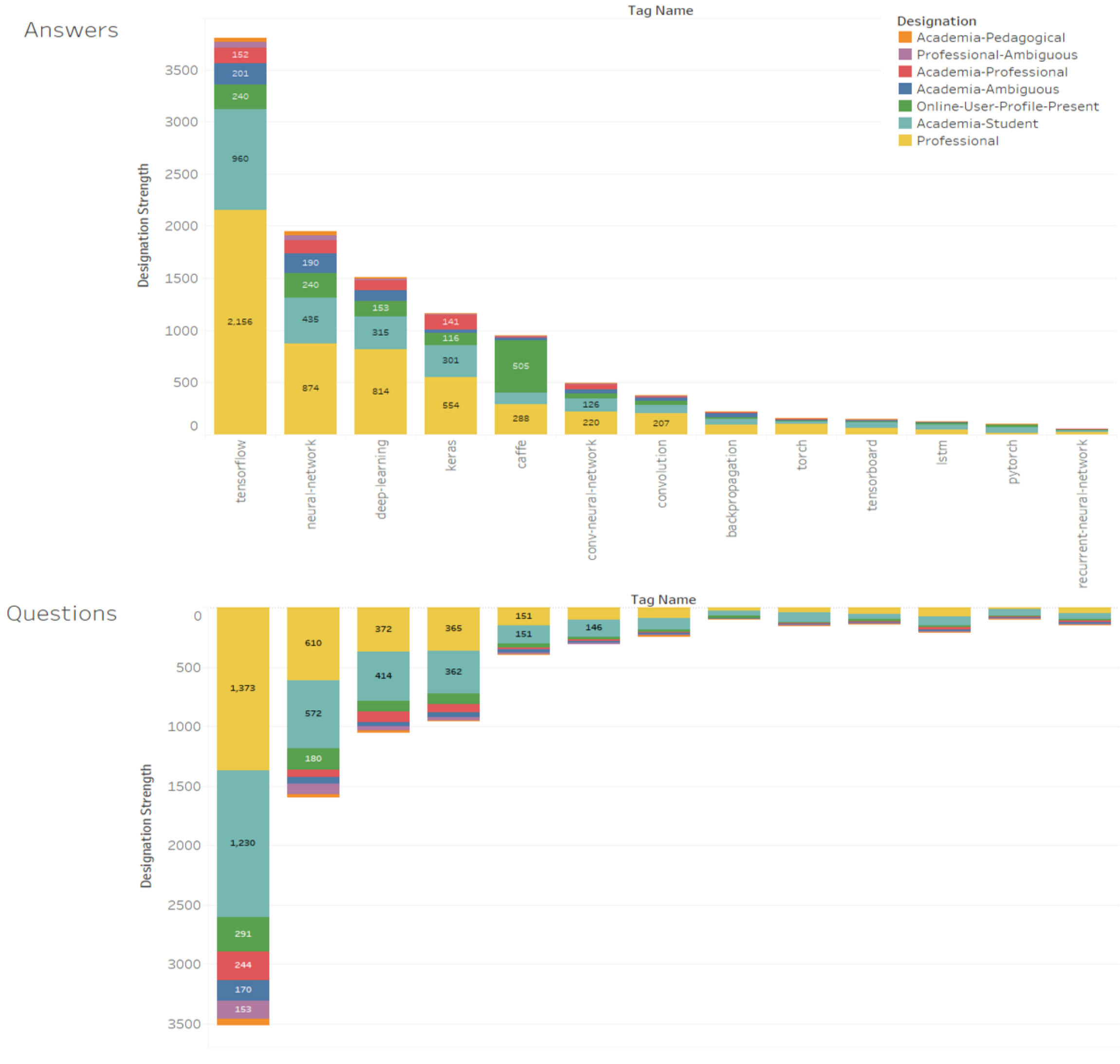
Using SWUDR range as a metric for categorizing users with questions and answers within a specific tag, we were able to observe a similar type of trend, i.e. that most of the questions and answers posts lie in the low score range and that they are being asked and answered respectively by users that belong to low SWUDR range. However, the difference from using reputation as a metric was that we were able to see more granular results in both questions and answers as we were able to observe better distribution within medium and high score range. The coarse level distribution was also preserved within each score range as well. This can be seen within tags like keras, caffe, lstm etc.

We also observed that more users with high SWUDR score are asking and answering questions and answers belonging to lower score ranges. This pattern could be more prominently exposed using the SWUDR rather than reputation as a metric and we can observe this pattern in tags like tensorflow etc. Another interesting discovery was that most of the tags have a same upper limit for SWUDR score. After performing further manual inspection we found that there existed a single user with high SWUDR score who had answered questions spanning across most of the tags used in our experiment. The tags encapsulating those questions which that user had answered had the same SWUDR high range. Also it is surprising to note that the same user has a low overall reputation but his/her SWUDR score existed in the high range for SWUDR which is specific to deep learning tags.

## 4.3 Research Question 2

*How long does a questioner have to wait before he/she can expect acceptable answers for his/her deep learning tagged questions?*

To address RQ2, as we described in the methodology, we grouped the processed data of the top 3 most upvoted an-



Sum of Designation Strength for each Tag Name. Color shows details about Designation. The view is filtered on Designation, which has multiple members selected.

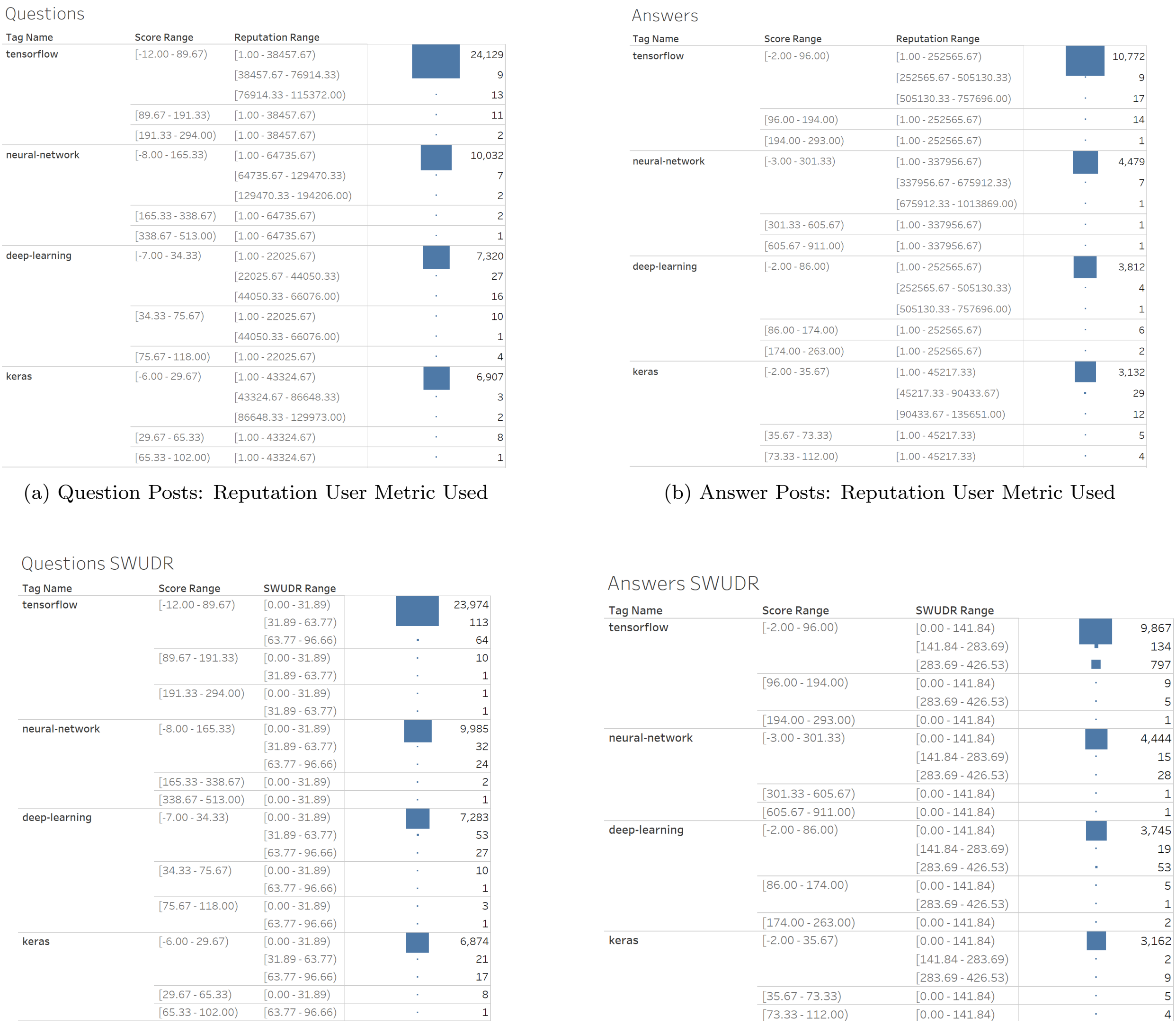
Figure 2: RQ1.1: Designation distribution amongst the different tags

swers for each question according to the tags which are associated with it and calculate the mean across the same rank of answers. The 1st, 2nd and 3rd order mean response times for the top 10 designations are displayed on Figure 4.

From the graph, we found an interesting trend present across 19 out of the 30 tags we used for the experiment. The 3rd most upvoted answer was always the quickest to appear amongst the top 3 most upvoted answers. Although this might seem trivial, it provides an interesting insight. The quickest emerging answer post is usually the one which has a great outreach to the viewers who have subscribed to the question initially and can be seen as the answer which is most experimented with as an initial solution. Users can use the result of RQ2 as a reference to gauge the 3rd order response time for their question and can use that as an upper bound to wait for an answer that can be viewed as an acceptable (although sub-optimal).

Amongst the tags we have, we can divide them roughly into 2 buckets, concept-based and library/package based. An example for the earlier can be conv-neural-network while an example for the latter can be tensorflow. We observe that for the library/package based posts, the 1st and 2nd order response times are quite near to each other. The answer seekers who prefer to rely on the "best" answer would have to wait to see the most upvoted answer develop into the best answer. During this period, the similar scored answers would compete for upvotes until the best answer emerges. The waiting period arises from the fact that the 1st and 2nd best answers are difficult to differentiate the mean response times of both are similar.

For the concept-based posts, we observe that the 2nd most upvoted answer usually comes later than the most upvoted answer. We hypothesize that this trend exists as a result of the concept-based nature of the posts. The answering community view the most upvoted answer as the answer most suitable for clarifying a question, but they may want



Questions

| Tag Name | Score Range | Reputation Range | |
|---|---|---|---|
| tensorflow | [-12.00 - 89.67) | [1.00 - 38457.67) | 24,129 |
| | | [38457.67 - 76914.33) | 9 |
| | | [76914.33 - 115372.00) | 13 |
| | [89.67 - 191.33) | [1.00 - 38457.67) | 11 |
| | [191.33 - 294.00) | [1.00 - 38457.67) | 2 |
| neural-network | [-8.00 - 165.33) | [1.00 - 64735.67) | 10,032 |
| | | [64735.67 - 129470.33) | 7 |
| | | [129470.33 - 194206.00) | 2 |
| | [165.33 - 338.67) | [1.00 - 64735.67) | 2 |
| | [338.67 - 513.00) | [1.00 - 64735.67) | 1 |
| deep-learning | [-7.00 - 34.33) | [1.00 - 22025.67) | 7,320 |
| | | [22025.67 - 44050.33) | 27 |
| | | [44050.33 - 66076.00) | 16 |
| | [34.33 - 75.67) | [1.00 - 22025.67) | 10 |
| | | [44050.33 - 66076.00) | 1 |
| | [75.67 - 118.00) | [1.00 - 22025.67) | 4 |
| keras | [-6.00 - 29.67) | [1.00 - 43324.67) | 6,907 |
| | | [43324.67 - 86648.33) | 3 |
| | | [86648.33 - 129973.00) | 2 |
| | [29.67 - 65.33) | [1.00 - 43324.67) | 8 |
| | [65.33 - 102.00) | [1.00 - 43324.67) | 1 |

(a) Question Posts: Reputation User Metric Used

Answers

| Tag Name | Score Range | Reputation Range | |
|---|---|---|---|
| tensorflow | [-2.00 - 96.00) | [1.00 - 252565.67) | 10,772 |
| | | [252565.67 - 505130.33) | 9 |
| | | [505130.33 - 757696.00) | 17 |
| | [96.00 - 194.00) | [1.00 - 252565.67) | 14 |
| | [194.00 - 293.00) | [1.00 - 252565.67) | 1 |
| neural-network | [-3.00 - 301.33) | [1.00 - 337956.67) | 4,479 |
| | | [337956.67 - 675912.33) | 7 |
| | | [675912.33 - 1013869.00) | 1 |
| | [301.33 - 605.67) | [1.00 - 337956.67) | 1 |
| | [605.67 - 911.00) | [1.00 - 337956.67) | 1 |
| deep-learning | [-2.00 - 86.00) | [1.00 - 252565.67) | 3,812 |
| | | [252565.67 - 505130.33) | 4 |
| | | [505130.33 - 757696.00) | 1 |
| | [86.00 - 174.00) | [1.00 - 252565.67) | 6 |
| | [174.00 - 263.00) | [1.00 - 252565.67) | 2 |
| keras | [-2.00 - 35.67) | [1.00 - 45217.33) | 3,132 |
| | | [45217.33 - 90433.67) | 29 |
| | | [90433.67 - 135651.00) | 12 |
| | [35.67 - 73.33) | [1.00 - 45217.33) | 5 |
| | [73.33 - 112.00) | [1.00 - 45217.33) | 4 |

(b) Answer Posts: Reputation User Metric Used

Questions SWUDR

| Tag Name | Score Range | SWUDR Range | |
|---|---|---|---|
| tensorflow | [-12.00 - 89.67) | [0.00 - 31.89) | 23,974 |
| | | [31.89 - 63.77) | 113 |
| | | [63.77 - 96.66) | 64 |
| | [89.67 - 191.33) | [0.00 - 31.89) | 10 |
| | | [31.89 - 63.77) | 1 |
| | [191.33 - 294.00) | [0.00 - 31.89) | 1 |
| | | [31.89 - 63.77) | 1 |
| neural-network | [-8.00 - 165.33) | [0.00 - 31.89) | 9,985 |
| | | [31.89 - 63.77) | 32 |
| | | [63.77 - 96.66) | 24 |
| | [165.33 - 338.67) | [0.00 - 31.89) | 2 |
| | [338.67 - 513.00) | [0.00 - 31.89) | 1 |
| deep-learning | [-7.00 - 34.33) | [0.00 - 31.89) | 7,283 |
| | | [31.89 - 63.77) | 53 |
| | | [63.77 - 96.66) | 27 |
| | [34.33 - 75.67) | [0.00 - 31.89) | 10 |
| | | [63.77 - 96.66) | 1 |
| | [75.67 - 118.00) | [0.00 - 31.89) | 3 |
| | | [63.77 - 96.66) | 1 |
| keras | [-6.00 - 29.67) | [0.00 - 31.89) | 6,874 |
| | | [31.89 - 63.77) | 21 |
| | | [63.77 - 96.66) | 17 |
| | [29.67 - 65.33) | [0.00 - 31.89) | 8 |
| | [65.33 - 102.00) | [63.77 - 96.66) | 1 |

(c) Question Posts: SWUDR User Metric Used

Answers SWUDR

| Tag Name | Score Range | SWUDR Range | |
|---|---|---|---|
| tensorflow | [-2.00 - 96.00) | [0.00 - 141.84) | 9,867 |
| | | [141.84 - 283.69) | 134 |
| | | [283.69 - 426.53) | 797 |
| | [96.00 - 194.00) | [0.00 - 141.84) | 9 |
| | | [283.69 - 426.53) | 5 |
| | [194.00 - 293.00) | [0.00 - 141.84) | 1 |
| neural-network | [-3.00 - 301.33) | [0.00 - 141.84) | 4,444 |
| | | [141.84 - 283.69) | 15 |
| | | [283.69 - 426.53) | 28 |
| | [301.33 - 605.67) | [0.00 - 141.84) | 1 |
| | [605.67 - 911.00) | [0.00 - 141.84) | 1 |
| deep-learning | [-2.00 - 86.00) | [0.00 - 141.84) | 3,745 |
| | | [141.84 - 283.69) | 19 |
| | | [283.69 - 426.53) | 53 |
| | [86.00 - 174.00) | [0.00 - 141.84) | 5 |
| | | [283.69 - 426.53) | 1 |
| | [174.00 - 263.00) | [0.00 - 141.84) | 2 |
| keras | [-2.00 - 35.67) | [0.00 - 141.84) | 3,162 |
| | | [141.84 - 283.69) | 2 |
| | | [283.69 - 426.53) | 9 |
| | [35.67 - 73.33) | [0.00 - 141.84) | 5 |
| | [73.33 - 112.00) | [0.00 - 141.84) | 4 |

(d) Answer Posts: SWUDR User Metric Used

Figure 3: RQ1.2: Question and Answer Posts Grouped on Score and Sub-grouped on Different User Metrics

to add supplementary information which would aid in the clarity or furnish important details which might have been missing in the 1st and 3rd (the previous 2nd) most upvoted answer.

We consider the result of RQ2 to be more important as it serves as a reference to the deep learning community on Stack Overflow on when can a questioner expect acceptable answers to his/her questions, to appear after he/she has asked a particular question associated with a certain set of tags. Although it is computed with a simple statistical formula in this study, we propose to extend our study in future to use a regression model to predict the mean response time according to a large set of features which include the designations of the users most active within the tag, overlapping of tags, extent of details in the question, presence of code within the post etc.

## 5. THREATS TO VALIDITY AND FUTURE WORK

We acknowledge the following threats to validity with our study.

- The study does not consider the exhaustive list of all tags associated with deep learning and tags overlapping with machine learning may influence the results.

- Due to the possibility of a post being associated with multiple tags, the number of posts represented may be a result of posts of overlapping tag sets, although the ratio would remain preserved.

- For RQ1.1, more than half the posts were authored by users whose designations could not be determined from the "AboutMe" field and hence are "Unknown". This can be delegated as part of future work where we could request Stack Overflow administrators to allow access to the actual designation of the user.

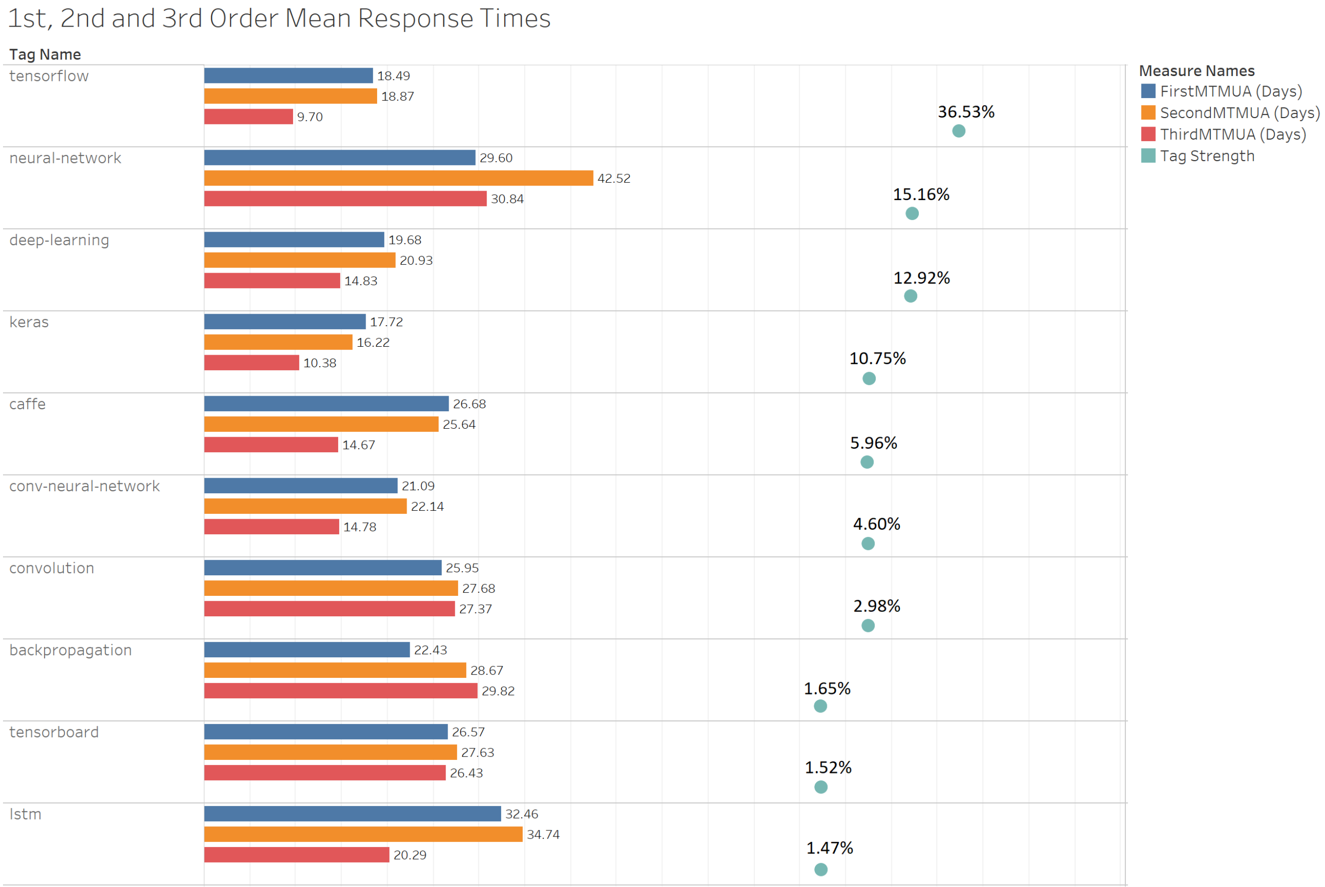


Figure 4: RQ2: 1st, 2nd and 3rd Order Mean Response Times Grouped by Tag

- For RQ1.1, the Designation to Keywords mapping was based on our understanding of each keyword and meticulous research was conducted on understanding the details of the responsibilities and actions encompassing the role represented by each keyword. Since this mapping is a result of manual judgment, other studies may differ in their versions of mapping and this possibility poses a threat to validity.

- It deserves to be mentioned that a potential threat to RQ1.2 and RQ2 is that the data used for processing answers does not include those posts from those users who did not have a valid "UserId". We would argue that including those answers would have posed a much more stronger threat to validity since any user metric such as Reputation & SWUDR would have been obsolete.

- The ranges which were used for RQ1.2 were very broad but were data driven. Future work would be to generate more granular ranges or derive the ranges from a clustering technique to drive inferences from the result.

- For RQ2, some tags such as "tensorflow", "keras", "neural-network", "deep-learning" etc. are very broad. The mean response times presented for those tags might not be a pragmatic representation of the actual response times.

## 6. CONCLUSION

According to the Gartner Hype Cycle July 2017 [1], Deep Learning is at the "Peak of Inflated Expectations" phase and deriving insights at this crucial period would benefit the research community greatly. In this paper, we hoped to maverick such an effort by attempting to answer 3 out of a large number of possible research questions which exist with the information available from Stack Overflow posts.

Despite the large number of "Unknown" designations, from RQ1.1, we could garner that the Professional community is highly active in both asking and answering deep learning questions and as expected, the Academia-Student community is next in the participation front. We are baffled as to why the Academic-Pedagogical community lacks participation, at least in the answer posts, since they would be best positioned to clarify conceptual questions at the least. The results from RQ1.2 were also unexpected since we expected more reputed users to be asking and answering high score questions instead of targeting the low score ones. Further research would perhaps yield interesting reasons about the result of RQ1.2. The simple study conducted for RQ2 yielded results which could be used by the deep learning Stack Overflow community as a reference of expected response times for acceptable answers to questions tagged with a set of deep learning tags.

We further elicited and acknowledged a set of threats to validity to our experiment and listed points of action that can be taken as future work to make our results more reliable and concrete. We hope to take this line of research forward at this crucial juncture of the Deep Learning technology by asking and answering actionable questions regarding the relevant activity stemming from the Stack Overflow community.

## 7. REFERENCES

[1] Gartner newsroom press release 2017. https://www.gartner.com/newsroom/id/3784363.

[2] H. Alharthi, D. Outioua, and O. Baysal. Predicting questions' scores on stack overflow. In *2016 IEEE/ACM 3rd International Workshop on CrowdSourcing in Software Engineering (CSI-SE)*, pages 1–7, May 2016.

[3] L. An, O. Mlouki, F. Khomh, and G. Antoniol. Stack overflow: A code laundering platform? In *2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 283–293, Feb 2017.

[4] A. Bacchelli. Mining challenge 2013: Stack overflow. In *The 10th Working Conference on Mining Software Repositories*, page to appear, 2013.

[5] A. Bacchelli, L. Ponzanelli, and M. Lanza. Harnessing stack overflow for the ide. In *2012 Third International Workshop on Recommendation Systems for Software Engineering (RSSE)*, pages 26–30, June 2012.

[6] D. Ford, A. Harkins, and C. Parnin. Someone like me: How does peer parity influence participation of women on stack overflow? In *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 239–243, Oct 2017.

[7] H. Hu, B. Liu, and P. Zhang. Several models and applications for deep learning. In *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, pages 524–530, Dec 2017.

[8] G. E. Lezina and A. M. Kuznetsov. Predict closed questions on stackoverflow. In *Proc. of the Ninth Spring Researcher's Colloquium on Database and Information Systems*, 2013.

[9] B. Lin and A. Serebrenik. Recognizing gender of stack overflow users. In *2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR)*, pages 425–429, May 2016.

[10] P. Morrison and E. Murphy-Hill. Is programming knowledge related to age? an exploration of stack overflow. In *2013 10th Working Conference on Mining Software Repositories (MSR)*, pages 69–72, May 2013.

[11] O. Odiete, T. Jain, I. Adaji, J. Vassileva, and R. Deters. Recommending programming languages by identifying skill gaps using analysis of experts. a study of stack overflow. In *UMAP '17 Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 159–164, July 2017.

[12] R. K. Saha, A. K. Saha, and D. E. Perry. Toward understanding the causes of unanswered questions in software information sites: a case study of stack overflow. In *ESEC/FSE 2013 Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*, pages 663–666, August 2013.

[13] J. Zhang, H. Jiang, Z. Ren, and X. Chen. Recommending apis for api related questions in stack overflow. *IEEE Access*, 6:6205–6219, 2018.