Why the vast majority of StackOverflow only posts one question

Rogier Slag
Delft University of Technology
Delft, the Netherlands
1507761
rogier.slag@gmail.com

Mike de Waard
Delft University of Technology
Delft, the Netherlands
4081064
mikedewaard@gmail.com

***Abstract*—While StackOverflow is a very well-known platform in the programming industry, the Q&A section is only effectively filled with content by a low number of users. Within that user group, around 75% only makes one contribution to the platform in total. Since the platform is that popular, how come the number of participants is that low?**

This paper defines a group called the *one-day flies* and tries to find reasons these users do not continue to contribute to the platform. By enabling users to become more active, the community itself could grow even further and continue to deliver answers to programmers all over the world.

To do this, we analyze a public StackOverflow database dump for several assumptions. These are compared between the one-day flies and the regular user group, in order to find statistical differences between these groups on certain level.

I. Introduction

StackOverflow is a Q&A community on the Internet, which is managed by the StackExchange platform. It focuses primarily on programming related questions, and is the biggest Q&A community on that specific topic.

StackOverflow has been gaining popularity during the last few years [1], and lots of research has been done on the data corpus of StackOverflow, such as [2], [3] and [4]. These and many other research papers on StackOverflow are about the users personalities [5], their field of expertise and on how these users contribute [6], but not so much research has gone into how the underlying structure of the StackOverflow community. With underlying structure we mean, for example the distribution of posts among users, the distribution of reputation among users, and so on.

This is why we started with some exploratory data analysis on the StackOverflow Data, which revealed that over 75% of the users posted only twice or even less during their whole registration period. Given this high amount of *one-day flies* we analyzed what happens to the content that these people contribute to the system. The initial idea was that most of these single-posts could be either marked as duplicate, very specific to a certain domain / programming language and do not get answered, or reveal a flaw in the documentation of an API. To elaborate a bit further on the flaw in API Documentation, a one-day fly can ask a question regarding how to use a specific API but not have any further questions for StackOverflow as he/she is an experienced developer. However during this research interesting findings regarding these initial beliefs emerged.

II. Related work

By being one of the largest, and most well-known Q&A platform regarding programming related questions, StackOverflow is an interesting community for researchers. Additionally StackOverflow actively releases parts of its database to the public, which proved to be of an extreme value for researchers. Therefore, it comes as no surprise StackOverflow has attracted a significant amount of researchers in the areas of software engineering, and social studies.

Several aspects of StackOverflow have already been investigated, such as how to build reputation [5], and how the reputation system encourages users to participate [6]. These works gave us insight in how the reputation system works and how this affects the user base. Other researchers [7]. investigated the deletion of questions on StackOverflow, which we used to verify some of our research questions.

A group of researchers from Canada [8] did work on the personality traits of StackOverflow users, which we used as base for the sentimental analysis for post and answers.

And finally related work was done in [9], where researchers looked specifically at the group of experts. Our work complements this work such that we look at the group that was left out in this research.

III. Exploratory Data Analysis

For this analysis, the StackOverflow data as provided by the 2014 MSR Challenge was used. This data was then preprocessed and saved to a MS-SQL server. This allowed for rapid processing, due to which large parts of the analysis could be done on the entire data set.

After this preprocessing step was completed an exploratory analysis was done on the data set. Since StackOverflow is a regular Q&A website, focused on programming languages, there was a suspicion it would adhere to general community dynamics specifics. An indicator of this would be a large number of inactive users, a small set of users which are more active, and a very small set of very active users.

To test this, clustering of several user groups was required. Even though this can be done using several aspects, the number of posts is for most communities applicable, as this is a direct measure for contributions to the community. Moreover StackOverflow has a special reputation system, which gamifies the platform. The reputation is acquired by users through the system and reflects how the contributions of people are received by the community. Using the reputation points earned by the users, they can be classified into several groups. Based on general community dynamics, a clustering in three groups seemed appropriate, as this would result in inactive (or one-day flies), active and very active user groups.

In the dataset we got from StackOverflow on september 26th in 2014, a total of 2.063.174 users had a reputation of one, whilst the total amount of users of StackOverflow is 3.473.095.This causes the median to equal to one, whereas the third quartile is just 13. Compared to the maximum value of 709.269 it is safe to conclude that a lot of users do not actively participate in the community, either by asking questions or proposing answers.

This can also be seen in the analysis of the number of posts a user made. 75% of the people has made a maximum of two posts. Again here the maximum is shifted towards the other end; the user with the most posts has acquired 30.043 posts on the website. It therefore seems a large majority of the users has no need nor wish to participate more in the community.

When the clustering into three groups is completed using k-means theory (see figure 3), three clusters are obtained, based on the reputation they achieved. The group of what we refer to as one-day-flies is the largest, with a total of 3.469.897 users. The medium active user group is significantly smaller and only contains 3.012 users with an average reputation of 38.000. The group with the most reputation is by far the smallest, and only contains 186 users. The average reputation of this group is then very high (almost 200.000).

When correlated with the number of posts or comments, it appears there is a linear relation between those on one hand and reputation on the other hand. The center groups of the clustering almost form one straight line when plotted.

IV. Initial Assumptions

Before conducting the actual research, a number of assumptions was made. These assumptions formed the base for the research questions we try to answer.

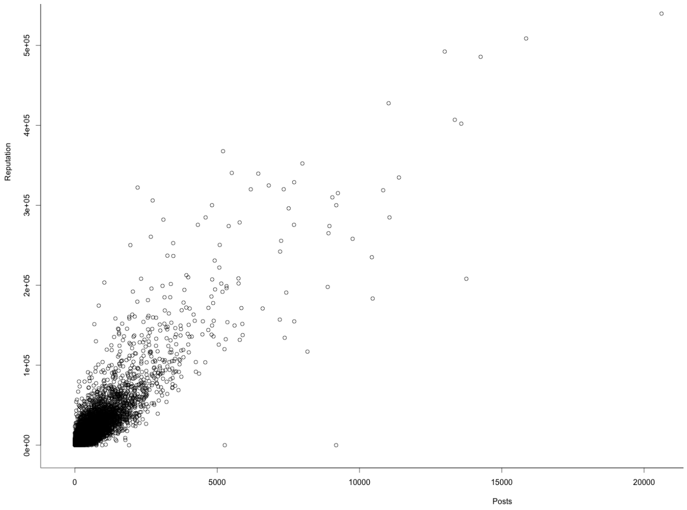

Fig. 1. The number of posts a user made compared to the reputation she gained

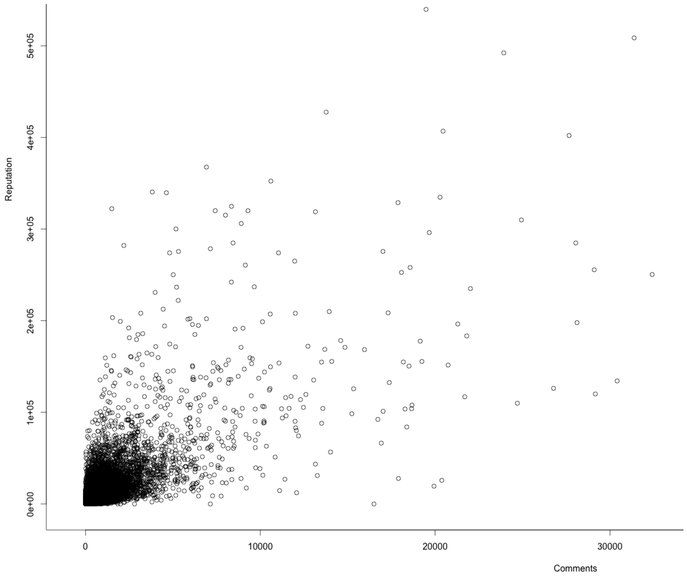

Fig. 2. The number of comments a user made compared to the reputation she gained

One of the earliest assumptions, which is not tested, was that the group with largest number of users which are not active on the site at all, consists of so-called *newbies*. These were more likely to not understand how the community of StackOverflow works, and may ask trivial questions which are not suitable for the level of expertise StackOverflow aims for.

More testable assumptions were also made. Many users which would register on the website could, because of their unfamiliarity with the system, ask questions which had been asked previously. These questions are quickly flagged by the community as *duplicate*. However this may be the appropriate action for the community as a whole, it can be extremely

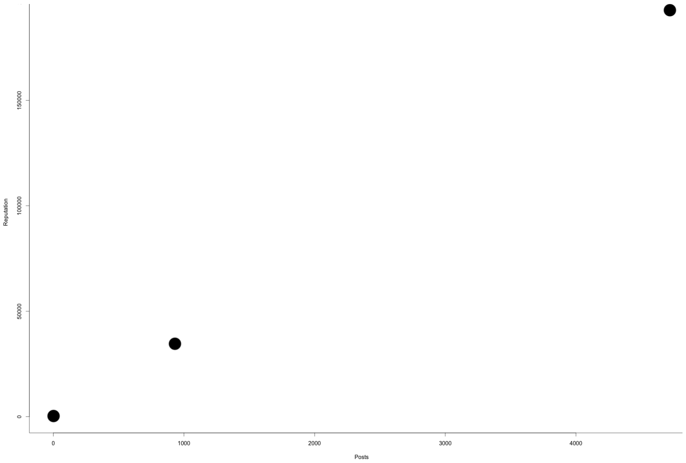

Fig. 3. The clustering of the user groups

discouraging for the user which posted the question. Our research question is then: *Are users which have posted only one question more likely to have created questions which are marked as duplicates*.

It may also be the case that these users got discouraged due to some other reason, such as asking questions about uncommon topics. In case their question was regarding a topic with hardly any experts on the StackOverflow platform, the user again can get discouraged by the lack of comments or answers to guide them into the right direction. However, simply parsing the content of the question is not sufficient. In order to be found by possible experts, it is also required for users to use the correct tags for their question. Otherwise, the question might simply get lost in enormous question set of StackOverflow. This gives the following research question on this topic: *Are users which have posted only one question more likely to have posted questions with uncommon tags*.

Another possible explanation for the large number of users with a low question count can also be found in deletions. Users themselves can delete questions after posting them, and so can the community's moderators. This can have quite an impact. As studied in [7], deleted questions usually are of a very low quality (the quality is considered low compared to the expectation of the StackOverflow community). Users unfamiliar with the ethos of StackOverflow may inadvertently post such questions, only later realizing these do not belong there, and be deleted (or be deleted by moderators). This gives rise to the following research question: *Are users which have posted only one questions more likely to have their posts removed (either themselves or by a moderator / community)*.

Additionally it may be the case the user question was not viewed as much as they had anticipated, resulting in fewer good answers. It may be the case their question was listed lower due to their lack of reputation, or simply got lost in the list of questions. Apart from that, it is also possible that questions from people with no post history simply get opened less often, resulting in lower quality answers. Since a low view count can also, to some extent, discourage the question askers, since he or she may feel neglected by the community. The research question following from the above yields the following: *Are users which have posted only one question more likely to attract less views to the question he or she posted*.

The final assumption that was made is that new users simply got let down by the community: they asked a question which did not violate any of the StackOverflow rules, but did not receive a (satisfactory) answer anyway. This would probably the most discouraging situation, since the user did not do anything wrong, but did not receive help due to aspects out of his or her control. That might cause them to lose the confidence in the community itself. This yields the following research question: *Are users which have posted only one question more likely to have received no answer to their original question*.

V. Methodology

In this section we describe the methodology used to gain answers on our research questions which we specified in the section IV.

A. Processing the data

The data used was made available by StackOverflow itself and also used for the 2014 MSRChallenge [10]. The export consisted of the (filtered) data of several key tables of the StackOverflow website. The data was formatted in XML files, which makes it hard to query directly on the data. This is why the decision was made to implement a multithreaded data loader which stored the XML data into a MSSQL database. Although the process took two days to complete, this proved to be worth the effort: it allowed to do large analysis on the entire data set instead of restricting the analysis to subsets.

For the actual querying, statistical analysis, and plotting of the data a mixture of Ruby, C#, and R was used. Whereas Ruby was used for mining additional data, both C# and R showed their value in querying the data and performing statistical analysis. In the end R was used to generate any plots which were able to shed more light on the data.

B. Evaluation of Duplicate questions

In order to answer the research question *Are users which have posted only one question more likely to have created questions which are marked as duplicates*, a comparison between percentage of duplicate questions between all users and one-day flies should give similar results. Thus the methodology is to first find all users with only one question asked, then check whether the question is marked as duplicate (by checking on LinkTypeId in the PostLinks table) and do the same for all questions in the StackOverflow data. This holds because the total amount of questions asked by one-day flies is less than 8% of the complete question set. After retrieving these results a simple comparison should show whether one day flies post more duplicate questions in comparison to the complete user base.

C. Evaluation of Uncommon tags

In [7] researchers were able to find over 3.300 tags that only occur in questions of quality far below StackOverflow standards. The research question *Are users which have posted only one question more likely to have posted questions with uncommon tags* is based on these findings, as similar results might be found. The methodology for answering of this research question is to look at top 50 most popular tags given to one-day fly questions in comparison to the top 50 tags given in all questions in Stack Overflow. This should give enough insight in the most popular tags and the existence of a popular uncommon topic among one-day flies.

D. Evaluation of deletions

The number of deletions is not directly obtainable from the database dump as provided by StackOverflow. Therefore two sample sets were taken, with a size of 150.000; one for people with a low post count, and one for people with a high post count. For each of these sets, the corresponding StackOverflow page was fetched, and its DOM structure was parsed. This allowed to extract whether the question was deleted, or possibly closed.

E. Evaluation of less views

It might be possible that one-day flies simply have received less views on their posts. In order to correctly determine this, data set was assembled with a distinction between one-day-flies and regular users. To examine the results, it might be useful to eliminate the highest viewed questions and the least viewed questions from the data set. For this, we select the percentages of 5%, 10%, and 20%. Based o the results with exclusion of these 5%, 10% and 20% we can answer the research question *Are which have posted only one question more likely to attract less views to the question he or she have posted.*

F. Evaluation of unanswered questions

As stated earlier, it may be possible for one-day-flies that their questions are simply not answered that often (especially compared to veteran community members). To model this, two sets were taken. The set containing the questions for the one-day-flies had a size of 1.016.696, whereas the set containing the questions for the more active users consisted of 6.578.818 questions.

VI. Results

In this section we present the results of the evaluation of each research question as described in section V

A. Duplicate questions

The number of duplicate questions is surprisingly low. One could expect that this would be high on a public Q&A community where a large percentage of the users is unfamiliar with platform. However, only 2.2% of the one-day flies questions is marked as a duplicate. Meanwhile, this percentage increases to 2.9% for the regular users. This would be surprising, since one can assume that regular users are more familiar with the platform and its search functionality. Based on this data we can say that the answer to the research question *Are users which have posted only one question more likely to have created questions which are marked as duplicates* is no.

B. Uncommon tags

Surprisingly enough the top-5 tags of both one-day flies and regular users are identical (although ordered different). Although there are a number of tags which are specific to one-day flies, a large number of them is very similar.This allows us to answer the research question *Are users which have posted only one question more likely to have posted questions with uncommon tags.* with no.

C. Deleted questions

As it turned out, the number of deleted questions for low post count users was 15.4%, whereas it was 10.9% for the high post count users. However, it cannot easily be determined what the reason for deletion was. As explained in [7], there are several reasons for a question to be deleted. Without additional information, more research is needed to determine whether this can be a factor in the user participation. This research question therefore can not yet be answered.

The same analysis also gave interesting insights in the number of closed posts. For both groups, this turned out to be surprisingly low: 0.92% for low post count users, and 0.76% for high post count users. Questions can be closed for various reasons, such as being a duplicate or for being off-topic. Closed questions may be improved or deleted later on. However, the difference between the number is quite low , which allows us to answer the research question *Do one-day flies more often have their questions removed?* with no.

D. View counts

As it turns out, both groups have an average view count of around 190 (for one-day-flies this was 198, for regular users 187). However, the view count cannot be used only to determine this, since the standard deviation for these groups was quite high (respectively 400 for one-day-flies, but over 1.000 for regular users). To compensate for these effects, the dataset was modified to only include the posts which fitted into the 10% to 90% range of the view counts, and the results were rerun.

Using this reduced data set, with the extremes removed, a similar result is obtained, with an average view count of 127 for one-day-flies and 113 for non-one-day-flies. The standard deviations dropped a lot as well, to only 100 respectively 93. When the analysis is repeated by eliminating the extreme 5% of 20%, similar results remain.

Therefore we conclude that the posts of one-day-flies are actually not viewed less than those of non-one-day-flies, since it any case the view count of the latter group is actually lower. The research question *Are users which have posted only one question more likely to attract less views to the question he or she have posted* is therefore no.

E. Unanswered questions

Based on the data there seems to a difference between the two user groups. The group of one-day-flies has a larger percentage of unanswered questions: 17% compared to just 10% for the group which has a higher post count. Here we see a difference between the two groups. It might very well be the case user get demotivated because of this, especially if they saw other users receive answers to their posts much more often. Since this difference on the complete dataset is only 7% we can answer the question *Are users which have posted only one question more likely to have received no answer to their original question* with yes, but can argue that this is only a very slight difference, thus does not indicate the reason why one becomes a one-day fly.

VII. Qualitative research

Since the results did not reveal a clear reason which explained the existence of one-day flies, we decided upon a small qualitative research to get new ideas and see if something was overlooked that can be found in the data itself rather easily. For this the authors manually scanned through 50 posts and their accepted answer, answering the following questions:

1) Does the question contain code?
2) Does the accepted answer contain code?
3) Does the question contain a clear explanation?
4) Was the answerer friendly or unfriendly in his answer?
5) Did the question asker put in enough effort (was the answer not easily found by using any search engine)?

After answering these questions the results for the sample where as follows:

1	20 questions contained code
2	19 accepted answers contained code
3	All the 50 questions contained an explanation
4	48 answerers where friendly, 2 where unfriendly
5	11 times the answer was easily found by a search engine, the other 39 times they tried to find a solution themselves

The results of 1 and 2 show that over half the questions and accepted answers did not contain code, even though the questions where programming related. Most of them where regarding versioning systems, or development methods (such as Agile or database systems). However it still might be interesting to look at the code example quality as this seemed relatively poor in the sample the authors took.

The result of 3 is not a surprise, however the results of 4 is interesting. Two answers where unfriendly, but still got accepted. This gives the idea for a new research question: Do one-day flies not come back due to negative feedback?

The result of 5 gives the idea that askers are relatively “lazy”. However, since the classification “lazy” is quite opinionated, this cannot be used as a direct research result. However the authors did find during this qualitative research was that a few questions where later answered by the question owner him/herself. This gave the idea that one-day flies might only use StackOverflow for self-documentation purposes.

To get a better insight in questions 1, 2 and 4: we did an automated search for this. For question 1 and 2 this was easily done using R. For question 4, a Ruby script with SentiStrength [11] was used to evaluate how friendly the answers and questions were on a scale of -4 (very unfriendly) to 4 (very friendly).

For question 1, it turned out that for non one-day flies 65% of the questions had code, and 58% of the answers (over the set of all questions with an accepted answer). For one-day flies these numbers were 70% and 65% respectively. This indicates one day flies give and receive a bit more code than the other users.

To answer the friendliness, a random subset of 10000 questions with an accepted answer was taken. It turned out that non one-day-flies questions had a niceness of 0.079, and the answers they received had a niceness of 0.059. This indicates that both the questions and answers were mostly formal and professional, and on average slightly positive.
Again this was a bit different for one-day flies: they had a question niceness of 0.196 and received an answer with niceness 0.092. Still the tone appears to be mostly formal and professional, but the underlying touch tends to be a bit more positive.

VIII. New Research Questions

Based on the work presented in the previous chapter, the authors formulated a number of new research questions. This might form the basis for future work regarding the participation of *one-day-flies* on Q&A communities such as StackOverflow.

A. Code example quality

In general it seems the quality of the code examples in questions of one-day-flies is lower compared to the quality of users with a higher reputation. One-day-flies generally include too much code (by simply pasting all the code into the questions code block. This discourages users to form an adequate explanation.

Apart from that, the general code quality itself is lower as well. Code is poorly modularized, which causes people to respond to those issues, before addressing the original problem of the question asker. He or she may then get irritated, since they receive feedback they simply are not ready for at that point in the learning process. The same goes for the people trying to figure out what the issue is: they need to wrap their mind around to some non-standard principles used by others, which are (from their perspective) less easy to reason about.

B. Negative feedback

As stated in the previous paragraph, it is often the case the StackOverflow community gives tips and advice to novices how certain problems should be structured. However, for the question asker this can be seen as negative feedback. It does not solve their problem, while it increases the work they have to do in order to solve the problem or even get community support.

Even without this feedback, the StackOverflow community has strict rules on the type of questions allowed on the platform. A quick search over time showed the authors that these rules were not strictly followed in the early days of the StackExchange platform, but are more followed these days. This causes many beginners questions, subjective discussions (*Is A better than B?*), or too-localized questions (*What does this regular expression do?*) to be voted for closing or deletion.

Since the community has answered a large number of questions, one can expect the same thing to happen for other things, such as *marked as duplicate*. The chances are increasing that somewhere in the database, a very similar question might have been asked already, and received an appropriate answer. Meanwhile, a beginner on the platform might not be aware of the exact terminology to find that specific question. Therefore they may consider their problem to be new and unique, causing them to post it. Several minutes later, they may find their question marked as a duplicate combined with a snide comment on ”how you can search on StackOverflow”. Little imagination is needed to see how a user would see this as a negative feedback from the community.

C. Self-answering

In section VII the authors found that some of the questions accepted answers where given by the authors of these questions. This could be done for documentation purposes, but also for attempting to gain reputation, while other answers might be better. However in order to explore this further, it is first interesting to evaluate whether one-day flies more often answer their own questions than non-one-day flies on StackOverflow.

A user does not get reputation for answering their own question (but they can get reputation if the answer is upvoted by others). Here there seems to be a clear trend: users have found and answer, and decided to share it with the community (even though they do not have a direct profit from it themselves). This seems to be an indication these users are not scared away from StackOverflow at that point (otherwise they likely would not have posted that answer).

D. Everything can be found

During the research the authors had several discussions with other developers regarding this high amount of one-day flies. In most of these sessions developers said to never ask a question on StackOverflow because the answer to their question is already available. This gave the authors a new idea that possibly almost all questions can be found, thus the chance that a user has two questions that are not answered on StackOverflow already is rather slim. This could cause the community to look like it has many one-day flies whereas the one-day flies just have one question that could not be found on StackOverflow before. If this is the case, then when taking into account the user growth, one should see a decrease in the amount of questions asked over time.

IX. Conclusion and future work

As indicated by the previous sections, there does not seem to be a clear reason within these statistics why the group of *one day flies* does contribute more to the platform. It therefore seems there are other reasons for one-day flies to stop contributing to the StackOverflow platform. In this paper, no definite conclusion can be given of the *why* the vast majority of StackOverflow only posts one question. However a number of research questions were evaluated with their statistics shown in figure 4 , and more possible open research questions which can be found in VIII, allowing for future work.

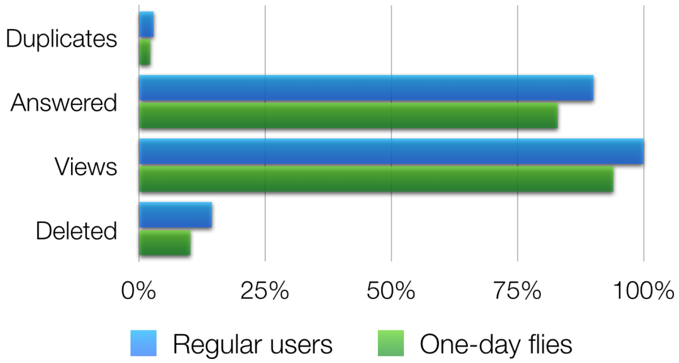

Fig. 4. Results of the analysis

References

[1] Anderson, Ashton, et al. *Discovering value from community activity on focused question answering sites: a case study of stack overflow.*, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.

[2] Treude, Christoph, Ohad Barzilay, and Margaret-Anne Storey. *How do programmers ask and answer questions on the web?*, Nier track.” Software Engineering (ICSE), 2011 33rd International Conference on. IEEE, 2011.

[3] Barua, Anton, Stephen W. Thomas, and Ahmed E. Hassan. *What are developers talking about? an analysis of topics and trends in stack overflow.*, Empirical Software Engineering 19.3 (2014): 619-654.

[4] Morrison, Patrick, and Emerson Murphy-Hill. *Is programming knowledge related to age? an exploration of stack overflow.*, Mining Software Repositories (MSR), 2013 10th IEEE Working Conference on. IEEE, 2013.

[5] Bosu, Amiangshu, et al. *Building reputation in stackoverflow: an empirical investigation.*, Proceedings of the 10th Working Conference on Mining Software Repositories. IEEE Press, 2013.

[6] Movshovitz-Attias, Dana, et al. *Analysis of the reputation system and user contributions on a question answering website: StackOverflow.*, Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ACM, 2013.

[7] Correa, Denzil, and Ashish Sureka. *Chaff from the wheat: characterization and modeling of deleted questions on stack overflow.*, Proceedings of the 23rd international conference on World wide web. International World Wide Web Conferences Steering Committee, 2014.

[8] Blerina Bazelli, Abram Hindle, and Eleni Stroulia *On the Personality Traits of StackOverflow Users*, 2013 IEEE International Conference on Software Maintenance

[9] Jie Yang, Ke Tao, Alessandro Bozzon, and Geert-Jan Houben *Sparrows and Owls: Characterisation of Expert Behaviour in StackOverflow*, 22nd International Conference, UMAP 2014, Aalborg, Denmark, July 7-11, 2014. Proceedings

[10] A. Ying, et al. *Mining Challenge - MSR 2015*, http://2015.msrconf.org/challenge.php

[11] Nielsen, Finn rup. *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs.* arXiv preprint arXiv:1103.2903 (2011).