Predicting Questions' Scores on Stack Overflow

Haifa Alharthi
School of Information Technology and Engineering
University of Ottawa
Ottawa, Canada
halha060@uottawa.ca

Djedjiga Outioua
School of Information Technology and Engineering
University of Ottawa
Ottawa, Canada
douti102@uottawa.ca

Olga Baysal
School of Computer Science
Carleton University
Ottawa, Canada
olga.baysal@carleton.ca

ABSTRACT

Developer support forums are becoming more popular than ever. Crowdsourced knowledge is an essential resource for many developers yet it can raise concerns about the quality of the shared content. Most existing research efforts address the quality of answers posted by Q&A community members. In this paper, we explore the quality of questions and propose a method of predicting the score of questions on Stack Overflow based on sixteen factors related to questions' format, content and interactions that occur in the post. We performed an extensive investigation to understand the relationship between the factors and the scores of questions. The multiple regression analysis shows that the question's length of the code, accepted answer score, number of tags and the count of views, comments and answers are statistically significantly associated with the scores of questions. Our findings can offer insights to community-based Q&A sites for improving the content of the shared knowledge.

Keywords

Crowdsourced knowledge; content quality; questions; prediction model; regression analysis.

1. INTRODUCTION

Nowadays, a great deal of information is published on the Internet every day. Many developers seek online help to solve problems they are facing. Questions and answers (Q&A) communities' goal is to provide an environment for such developers. A leading Q&A network is Stack Exchange, which consists of 149 Q&A websites. Stack Overflow is part of this network that attracts developers who are looking for programming-related solutions. Ensuring the quality of questions is essential in Q&A communities. Previous research found a strong positive correlation between the goodness of questions and the goodness of answers [6]. A question of high quality is assumed to attract more visitors (including experts) and more answers in short time. It also anticipated that a good question motivates answerers to do their best to receive more votes that increase their reputation points. In the long run, the activity and popularity of a Q&A community increases if the quality of the shared content (i.e., questions and answers) is high.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CSI-SE'16, May 16 2016, Austin, TX, USA

© 2016 ACM. ISBN 978-1-4503-4158-5/16/05…$15.00

DOI: http://dx.doi.org/10.1145/2897659.2897661

1.1 Stack Overflow

Stack Overflow receives more than six thousand questions on a daily basis [5]. Any registered user in Stack Overflow can post a question and other users can provide answers. Users could give up or down votes for questions and answers. Upvotes indicate that users find the post helpful, well-researched and thought-provoking, whereas downvotes mean that users may think the post lacks real explanation, has false information or ill-researched[1]. Each question and answer has a score that makes up the sum of up and down votes [6]. Users are encouraged to ask, answer and improve quality of the posts as they gain their reputation points by any upvotes their post receives. Many privileges go hand in hand with a high reputation such as voting to, commenting on and editing others' posts. The author of a question can mark an answer as "accepted", which denotes that the response meets the author's needs. An accepted answer does not necessarily imply that it is the best response for the community. Furthermore, a question must be associated with at least one tag and at most five tags. A tag describes the topic of a post, e.g., a programming language. When the visitors of the site selects one tag, they can access all the questions related to the tag's topic. Thus, tagging may increase the number of the question's readers who are interested in the topic[2].

To our knowledge, Stack Overflow employs certain procedures to ensure high quality of questions. To avoid having redundant questions, the website offers a list of similar posts to users before submitting their questions. The website also allows users to modify their posts or posts of others in order to make posts more readable and understandable. In addition, users can add comments to make clarifications or updates on their posts. However, more factors play a role in enhancing the quality of questions.

1.2 Problem Statement

In this paper, we conduct a quantitative study to investigate what factors impact the quality of questions. We examine many features related to the content and formatting of a post. These factors include the number of code blocks, links, tags, and paragraphs, as well as the length of the title, code and question body. Other factors related to the con-

[1] http://stackoverflow.com/help/why-vote
[2] http://stackoverflow.com/tour

tent are the polarity and subjectivity of a question. These factors can help in predicting the score of a question even before it is submitted and exposed to the community. If the approach was adopted, the website could guide users to enhance the content or the formatting of their questions prior their posts being published. We also consider another set of factors about the number of interactions associated with the question. These factors include the number of answers, favourites, views, comments, as well as the score of and the time to the accepted answer. These features may assist in the prediction of questions to be deleted or closed. For instance, inadequate posts or spam need to be early detected and deleted by the community.

Our research question investigates factors that can influence the score of a question on Stack Overflow. Previous research [6] has shown that the average of answers' scores on Stack Overflow is a good predictor of the score of questions. It also considered other factors such as questioner's reputation, the length of the question and the title and the number of answers, favourites and comments within the first 24 hours. In our research, we are adopting the same factors without a time frame, mainly because we are interested in an overall quality measurement.

In summary, our work makes the following contributions:

- To our knowledge, many factors investigated in this study have not been studied in previous research work. They include the ratio of post length to the number of paragraphs, a question's polarity and subjectivity values, accepted answer's score and the time to accepted answer. These features are explained in details in Section 3.2.
- An extensive analysis was conducted to study the relationship between the score of questions and the selected factors including a multiple linear regression model that was developed to predict the score of a question.

The rest of the paper is organized as follows. Section 2 discusses relevant research on prediction of the quality of questions and answers, as well as prediction of posts' closure and deletion. Section 3 describes our methodology including the dataset used, selection of factors and data analysis. Section 4 describes the results of the study. Section 5 discusses implications and threats to validity. Section 6 summaries our main findings and suggests possible future directions.

2. RELATED WORK

We now describe relevant research on the topic of prediction of the quality of questions and answers, as well as prediction of the question to be closed or deleted.

2.1 Prediction of the Quality of Questions and Answers

Many researchers have addressed the problem of quality prediction of online content including questions and answers. Yao et al. [6] investigated the quality of questions on Stack Overflow, as well as the quality of answers. For a question or answer to have high quality, it should have a high score. After computing Pearson correlation between the quality of questions and the average of their answers, a strong relationship was found. Features used for question prediction are questioner's reputation and the number of the previous questions, the length of the question and the title, and the number of answers, favourites and comments received within 24 hours from the post creation time. Features used for the answer quality prediction are the answerer's reputation and the number of previous answers, the answer length, and the number of comments received within 24 hours. The baseline models are two separate linear regression models; one for predicting questions and the other for answers. These baselines were compared against the co-prediction approach. The latter predicts the quality of questions and answers together. When scores are predicted regression is applied, whereas classification algorithms are used to deal with classes (e.g., high vs. low). In the co-prediction task, the estimated score for an answer can be used to predict a question score and vice versa. The co-prediction methods surpass the baseline.

Agichtein et al. [1] studied the quality of content on Yahoo! Answers which is a Q&A community. Questions and answers were classified based on their quality. The binary classifier is trained to distinguish between a high-quality post and other posts. The features used are in three categories: content, e.g., punctuation and typos, usage, e.g., clicks and relationships between users which were modelled as a graph. The classifier could successfully detect a good question and answer.

Jeon at al. [4] focused on enhancing the retrieval performance of documents of Q&A communities. To retrieve relevant and good quality questions and answers, their quality is estimated first. Instead of relying on text-based elements to improve the retrieval system, non-textual features were employed. They include the number of times a post was printed, copied, recommended or clicked. In addition, information related to the user is taken into account. Pairs of questions and answers were annotated according to their relevance to a query. A pair can have good, medium and bad relevance tag. Pearson's correlation was computed between these relevance levels and the non-textual features. The answerer information, as well as the answer length had the highest correlation coefficient. The experiments also showed that the retrieval accuracy has improved after the quality of posts was measured.

2.2 Prediction of the Questions Closure and Deletion

Some research efforts investigated poor quality questions on Stack Overflow. These questions need to be detected in order to be removed or closed. Correa and Sureka [2] studied what the elements that help in spotting questions should be removed. To learn the features that deleted questions have in common, they isolated and studied the posts deleted in a five-year period. Their analysis included the structure of questions, the ways the community vote and the behaviour pattern of owners when deleting their own posts. Their research went further by employing machine learning algorithms to predict which question will be erased not far after it being published. The features used for prediction were in four categories: post content, owner profile, post style and features generated by the community. The prediction model has an accuracy of 66%. It was observed that once a post receives votes, usually long after its creation, the community starts voting it down. Moreover, it was demonstrated that questions are deleted by their owners in an attempt to preserve their reputations. Some high-quality questions are

undeleted right after unintentional deletion.

The work of Lezina and Kuznetsov [5] aims to predict questions that will be closed. The authors use machine learning algorithms including Random Forest and Support Vector Machine in addition to Vowpal Wabbit. The algorithms were compared to choose the classifier with the highest results. The text of questions was represented in two forms: vectors of `tf-idf` weights and topics identified by the Latent Dirichlet Allocation (LDA). In addition to the textual representation of question, many features related to users, posts and tags are employed. User features include reputation, close-votes. Features about posts are the number of code blocks, the number of links, the number of occurrences of dates and time, the number of digit. Another feature is the number of times each tag was associated with closed questions. To find the significant features in predicting closed questions, feature selection was conducted. The most important three features were found to be: time between opening the question and closing it, code blocks number, and a question's score. The least important is the number of links in a post. The algorithm attained that the highest results was Vowpal Wabbit when it worked only on 200 topics modelled by LDA.

3. METHODOLOGY

In this section, we present a description of the Stack Overflow dataset and the extraction process, an explanation of how the independent variables were selected and computed, and how we performed our data analysis.

3.1 Dataset Description

In this study, we used a data dump of the Stack Overflow content which was published from 2009 to 2014. We obtained this dataset from the MSR Challenge 2015[3] [7]. The dataset consists of eight XML files from which we used only two: "posts" and "users". After consulting the schema of the dataset[4], we found that the two files contain all the attributes we need. We then mined the files of large sizes (1G and 28G). First, the XML files were parsed using lxml, a Python library. Then, 12,077 questions were extracted using an iterative reading process. For a question to be retrieved, it must have all the attributes filled and must not have been edited, closed or owned by the community. Edited questions are filtered out because their score may be different before and after modification. Closed questions are either redundant, off-topic, ambiguous, too broad or based on personal opinions[5]. They were not included in our study because scores of closed and unclosed questions might be influenced by different factors. Community-owned posts are excluded because they are based on the collaboration of a group of users, and the computation of up and down votes is not similar to regular posts[6]. Then, for each question, the ID number of its accepted answer was used to retrieve its record from "posts.xml". Similarly, based on the question owner's ID, the user record was retrieved from "users.xml". The records of accepted answers and owners were mapped to their questions.

[3]http:2015.msrconf.orgchallenge.php#challenge_data
[4]http://meta.stackexchange.com/questions/2677/database-schema-documentation-for-the-public-data-dump-and-sede
[5]http://stackoverflow.com/help/why-vote
[6]http://stackoverflow.com/help/closed-questions

3.2 Factors Determination

In this work, we considered some variables that are commonly used in the literature of questions and answers quality prediction. Similarly to Yao et al. [6], we used questioner's reputation, the length of the question and the title and the number of answers, favourites and comments. Other features used by Lezina and Kuznetsov [5] to predict closed questions were also considered here. They include the number of questions' views, code blocks, links and the length of the code.

We also tried additional variables that were thought to be worth examining. We believe that a question's number of tags, the score of the accepted answer, time to accepted answer, the ratio of body length to paragraphs, polarity and subjectivity value may play a role in the determining its quality. The greater the number of tags, the greater the number of potential users who may visit the post. Also, we cannot completely analyze the quality of a question without taking into account its accepted answer. That is why we filtered out any question that does not have an accepted answer. A question with good quality is expected to receive a quick and high-quality accepted answer due to the question clarity. We also considered the ratio of body length to paragraphs as lengthy posts are more readable if they contain a corresponding number of paragraphs. Previous research [5] studied the number of sentences that start with "I" and "you". Instead, we used the subjectivity value that indicates if the text is more objective or subjective. The polarity value of a question show if it is positive, neutral or negative. Some of the factors above are available in the extracted data, and the others need to be computed as follow:

- The length of the body of a question is considered as its number of characters after the deletion of code blocks, links and HTML tags (or other tags).
- The length of a title is its number of characters.
- The number of code blocks in a question was calculated using the tag `<code>` which precedes any code block.
- The number of links in a question was also computed by the tag `</a>`.
- The number of tags is the count of the tags associated with the question.
- The length of the code is the number of characters in all code blocks in a question.
- The time to accepted answer is the time difference between the creation of the accepted answer and its question. Since we noticed that the time is very short, we used minutes to represent the time gap.
- Polarity value is computed by a public Python library, Textblob . The value can be in a range from -1.0 to 1.0. A question with zero polarity value is neutral. A value greater than zero denotes a positive question, while a value smaller than zero means a negative question.
- Subjectivity value was computed using the same library. The subjectivity of a question ranges from zero to one, where zero means the question is highly objective, and one says it is highly subjective. Before computing the polarity and subjectivity value of a question, it is cleaned from code, links, and tags.

- The ratio of body length to paragraphs was calculated as the body length divided by the number of paragraph tags plus one.

We used Python scripts to perform calculations and cleanup. The final dataset consists of 12,077 questions with creation date between August 2008 and March 2009. Our dataset contains no duplicates or missing values since posts were extracted only if they have complete data.

3.3 Data Analysis

We performed an empirical analysis of the effect of sixteen numeric independent variables on the dependent variable, the score of questions. The scores of questions extend from -3 to 1,055 with an average of 12 and a median of 5. Figure 1 illustrates the distribution of scores of questions. We noticed that the dependent variable is skewed since 6,223 of the posts have a score of five or less. We executed a normality test, Kolmogorov-Smirnov, on all the variables and found that they are not normally distributed with significance equals to .000 (as reported by SPSS). Hence, we applied nonparametric tests in the rest of the paper.

4. RESULTS

In this section, we present the results of our data analysis. First, we measured a correlation between the independent and dependent variables. For variables with low correlation coefficients, more analysis is performed to understand their relationship with the scores of questions. Regression analysis is also carried out.

4.1 Spearman's Rank Correlation

A Spearman's rank correlation was conducted to test the dependency between the factors and the scores of questions. As Figures 2, 3, 4, and 5 demonstrate, four variables have a high correlation with the dependent variable namely views count (r=0.658), answers count (r=0.459), accepted answer score (r=0.777), and favourite count (r=0.707), where p-value is lower than 0.05. Another set of variables has a correlation coefficient greater than 0.1 at significant level equals 0.01. They are comments count (r=0.161), body length (r=-0.133), reputation of owner (r=0.132), and time to accepted answer (r=-0.131), p-value is lower than 0.05. The rest of the variables have a correlation of zero or slightly higher.

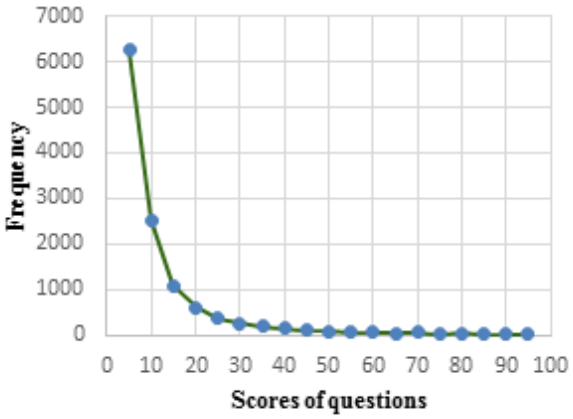

Figure 1: The distribution of the scores of questions.

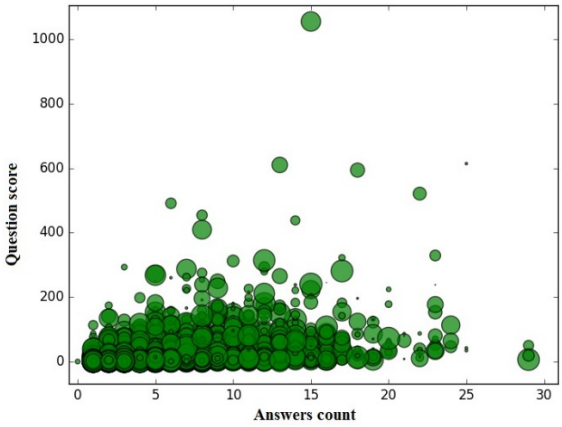

Figure 2: Correlation between the scores of questions and the number of answers.

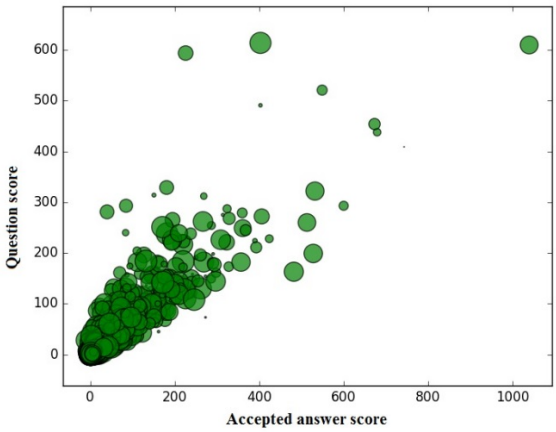

Figure 3: Correlation between the scores of questions and their accepted answers.

4.2 Extensive Analysis

More analysis was done using SPSS to understand the relationship between the dependent variables and the factors with low correlation coefficients. To check if the results are influenced by outliers, we removed any question with a score less than zero or more than 500. This did not show any improvement. We also deleted 7,701 records that have code blocks equals to zero, but the Spearman's correlation coefficient decreased from 0.023 to 0.001. Similarly, questions that have less than one link were deleted, but the correlation between the number of links and the scores of the remaining 1,616 questions only improved the results slightly (r=.044). Moreover, the number of questions with no comments is 9,338. After removing them, we noticed that the correlation coefficient dropped from .161 to .052. We also thought the ratio of body length to paragraphs might have an influence on the scores only when questions are long because they become unreadable without proportionate paragraphs. Therefore, we deleted all questions with body length less than the median (389) and ran the Spearman's correlation again. The results were worse than before. More analysis was held in the following sections:

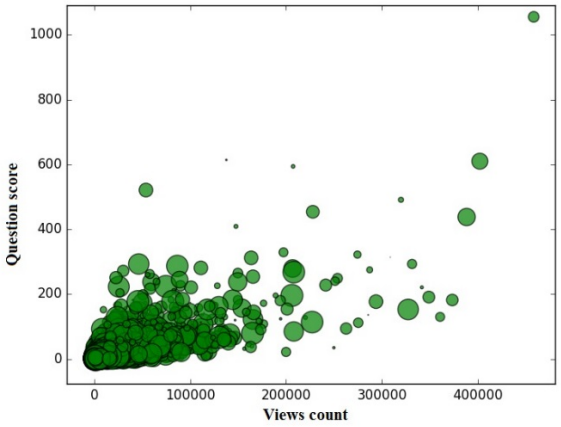

Figure 4: Correlation between the scores of questions and the number of views.

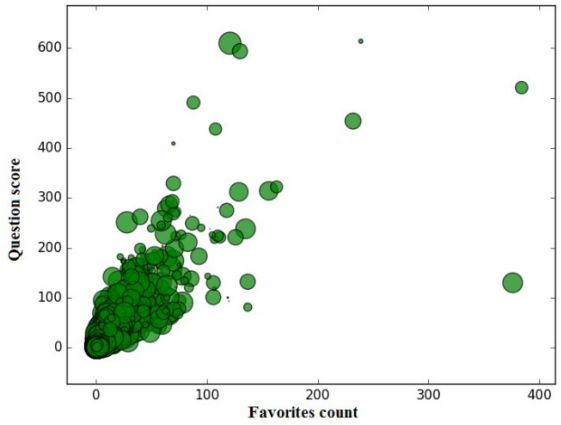

Figure 5: Correlation between scores of questions and the number of favourites.

4.2.1 Length of Code, Body and Title

It is intuitive to think that the size of the body and code have an effect on the scores of the question. A very long question or code is hard to read and require a great effort from the reader. A long title is also thought to be not appealing, and may lead to having few post's visitors. Hence, we studied them further.

Because the median of code length was zero, we removed 7,701 questions with zero code blocks. As a result, the correlation coefficient of code length increased from -.004 to -.174. Moreover, we divided the remaining 4,376 records into four quartiles of code length. Table 1 shows each quarter with its range of code length (in characters), average and median values of scores. One can see an upward trend in the scores as the code length gets smaller. For example, the first quarter has the questions with the shortest code (14–108) and also have the highest scores' average (18) and median (7). To test the difference between the quartiles, we applied the Kruskal-Wallis test. The null hypothesis of the test is that the distribution of scores is the same across the four categories of code length. This hypothesis was rejected at p-value equals to zero indicating that there is a significant effect of code length on the scores of questions.

Table 1: Descriptive statistics of the code length.

Category	Code Length	Average Score	Median Score
1st quarter	14–108	18.17	7
2nd quarter	109–227	14.92	6
3rd quarter	228–477	12.13	5
4th quarter	478–6,324	8.23	4

Table 2: Descriptive statistics of the body length.

Category	Body Length	Average Score	Median Score
1st quarter	4–238	16.92	7
2nd quarter	239–389	12.61	5
3rd quarter	390–614	10.63	5
4th quarter	615–5,666	8.41	4

Furthermore, the body length was also split into quartiles as illustrated in Table 2. As the questions' length increase, the average and median of their scores decrease. The questions in the first quarter have the smallest length (4–238 characters) and scores of 17 on average and median of 7, while lengthy questions (615–5,666 characters) in the fourth quarter have a lower average (8.4) and median (4). After applying the Kruskal-Wallis test, we found that the distribution of scores is the statistically different across the quartiles of body length at significance equals to zero.

Table 3: Descriptive statistics of the title length.

Category	Title Length	Average Score	Median Score
1st quarter	10–37	12.70	5
2nd quarter	38–49	13.27	6
3rd quarter	50–63	12.25	5
4th quarter	64–177	10.40	5

Similar to the length of code and body, we divided the title length into quartiles. The quartiles' range, the average of scores and median of scores are presented in Table 3. It is noticeable that the scores of the second quarter have the highest average and median. This quarter has titles of 38 to 49 characters which means they are of medium sizes. The first and third quarters have same average and median of scores. The last quarter that contains the longest titles has the least average and similar median. When the Kruskal-Wallis test was conducted, the null hypothesis was rejected at p-value equals to zero, which implies that there is a statistical difference in scores among the quartiles.

4.2.2 Reputation of Owner

Users on Stack Overflow gain upvotes when they provide good questions or answers. The total of up and down votes makes up the reputation. A high reputation of a user indicates that the questions and answers authored by her are of high quality. Thus, it is natural to think that an excellent reputation has a relationship with the score of questions. Yet, the two variables were weakly correlated with coefficient equals to 0.132. We created four categories based on

quartiles of the reputation. As seen in Table 4, the average and median of scores rise with the increase in the reputation of the owner. The Kruskal-Wallis test has also shown that the scores in the four categories come from statistically different distributions at p-value=0.

Table 4: Descriptive statistics of the user reputation.

Category	User Reputation	Average Score	Median Score
1st quarter	3–1,389	8.21	4
2nd quarter	1,390–4,414	11.91	5
3rd quarter	4,415–12,557	12.46	5
4th quarter	12,558–539,668	16.02	6

4.2.3 Time to Accepted Answer

We assume that a good question receives feedback soon because it may attract users to visit it and provide accepted answers in short time. In our sample of Stack Overflow data, the average and median of time until a question receives its accepted answer are 171 and 22 minutes correspondingly. When the time to accepted answer was split into quartiles as in Table 5, one can see that in categories where the time is short the average and median scores are significant and vice versa. The quartiles were shown to be statistically different from each other at the significance of zero when the Kruskal-Wallis test was applied.

Table 5: Descriptive statistics of the time to accepted answer.

Category	Time To Accepted Answer	Average Score	Median Score
1st quarter	0–7	16.07	6
2nd quarter	8–22	12.06	5
3rd quarter	23–132	10.69	5
4th quarter	133–1,439	9.47	4

4.2.4 Number of Tags

Users can annotate their questions with tags, which are keywords that describe the content of the question so that they attracts readers of interest. Naturally, one would think that a question tagged with more tags receives visitors and votes more than other questions. Since, the number of tags and scores of questions were weakly correlated, we performed more analysis. Tag numbers are already in the range from one to five so we considered them as categories of tags. Table 6 shows each category with its number of questions and scores' average and median values. The number of questions is not evenly distributed among categories as the questions annotated with two and three tags are greater than the total of the rest. The results are the opposite of our expectations. The median of scores is the same for all categories while their average decrease with the expansion of the number of tags. The scores of the four categories are statically different at p-value equals zero when tested by Kruskal-Wallis.

4.2.5 Polarity and Subjectivity

We assume that in Q&A community, questions are more negative because generally users seek help when they face a problem. Also, we expect users to be subjective since they mostly describe the problem from their perspective. However, the sample of data shows the opposite since about 70% of questions are positive and more than 60% are objective. Questions were divided based on their polarity and subjectivity. For polarity, questions with a value above zero are positive, less than zero are negative while zero values are discarded. For subjectivity, questions with a value greater than or equal to 0.5 are considered subjective and otherwise are considered objective. As Table 7 presents, we found that scores of positive and negative questions have almost the same average and median. The same observed when comparing objective and subjective questions in Table 8. We also used Kruskal-Wallis on both samples, but no significant differences were found.

Table 6: Categories of tags with their number, scores' average and median

Number of Tags	Number of Questions	Average Score	Median Score
One	1,334	13.01	5
Two	3,542	13.03	5
Three	3,863	12.16	5
Four	2,309	11.29	5
Five	1,029	9.89	5

Table 7: Polarity of questions.

Polarity	Average Score	Median Score	Number of Questions
Negative	11.65	5	2,054
Positive	11.52	5	8,121

Table 8: Subjectivity of questions.

Subjectivity	Average Score	Median Score	Number of Questions
Objective	12.11	5	7,716
Subjective	12.22	5	4,361

4.3 Multiple Linear Regression

To examine if the factors are powerful predictors of the scores of questions, we performed multiple linear regression using a Python library, Statsmodel. Multiple linear regression models try to predict the dependent variable Y by knowing the values of the independent variables x1, x2, x3...etc. [3]. The coefficient of determination (R-squared), which determines how well the regression line fits the data, of our regression model equals 0.879. This says that our model has a high predictive power. The Prob(F) is equal to zero which means the probability that the independent variables have no effect on the scores of questions is zero. Seven independent variables showed statistically significant association with the scores of questions at p-value= 0.05. The list of variables with their regression coefficient is as follows: answers count (0.2819), accepted answer score (0.3985), views count (0.0002), favourites count (0.9347), code length (-0.0005), comments count (0.5161) and tags number (-0.1891).

5. DISCUSSION

In this work, we explore the topic of the question quality prediction. This study investigates sixteen factors that may affect the quality of questions on Stack Overflow. However, many other elements would lead users to upvote a question. Users might like a question because they are interested in the topic, the question discusses a common problem, it is supported by facts, details and examples or it simplifies a complex issue. In our research, we focused on factors that are measurable based on the available dataset.

Our analysis shows that there are two independent variables, which are related to the format and the content, have a statistically significant association with the dependent variable. They are code length and tags number. Both correlate negatively with the scores of questions. This means the higher the scores, the fewer the tags which does not contradict with our belief that a greater number of tags makes a question more visible to the interested answerers. However, in Section 4.2.4, when comparing the distributions of questions of a different number of tags, the category of five tags had the least average of scores. The average might be biased because of the uneven number of questions in each category. Moreover, our results demonstrate that good quality questions have less code. Yet, the number of code blocks has no effect on the quality of questions. Therefore, many small blocks do not lead to a low question score but a long code does.

Five other powerful predictors of the questions scores are related to the interactions in the post. Interactions mean different actions applied to a question by the community such as asking for clarification and giving feedback. These predictors are views count, comments count, answers count, favourites count, and accepted answer score. Instead of calculating the average of all answers' scores [6], we focused on the accepted answer. We think that the accepted answer reflects the quality of a question more than other answers because if the question owners could write a well written and researched question, then they will not accept a poorly written answers. In the regression model, accepted answer's score is positively associated a question's score. Moreover, comments help in making clarifications or corrections to a question. We believe that the number of comments indicates the interest level of the post's visitors in understanding the question. This is one explanation of why they play a role in predicting questions' scores. Additionally, users favourite a question when it is of high quality so they would visit it again. It is reasonable that a high positive correlation exists between favourites and questions' scores. In fact, favourites are just another indicator of the quality of the question. Also, for a question to receive many upvotes, a high number of views is expected which explains the positive correlation.

5.1 Threats to Validity

In general, in the sentiment analysis including polarity and subjectivity, punctuations are important to analyze the text. In the body of the questions, there were many HTML tags such as the new line and new paragraph tags. The deletion of these tags may have affected the calculation of the polarity and subjectivity values. To evaluate the effect of the tags removal on the sentiment analysis calculation, we found the polarity and subjectivity values of questions with tags in them and the values have not changed at all or just slightly.

Moreover, some selected independent variables occur after the submission of the question. They are views, favourites, answers and comments. While the score of questions are changing, theses variables are changing in same time. We are not sure if the high scores lead to increase of these variables values. In other words, we cannot confirm if these variables are independent from the scores of questions.

6. CONCLUSION

This paper addresses the issue of questions' scores prediction on Stack Overflow. There are plenty of factors that might influence the scores from which we selected sixteen variables. We performed an extensive investigation to understand the relationship between the variables. The Spearman's correlation showed that the number of views, answers and favourites, as well as the accepted answer score have a high correlation with the dependent variable. When we divided the length of body, code and title, the reputation of owner and time to accepted answer into four categories, we found that the scores of the four categories are statistically different. Our regression analysis demonstrates that the chosen independent variables have a significant explanatory power.

There is a big room for improving the prediction of questions quality. The topics of tags may help the prediction of questions. One also can perform analysis on the writing style of the author in addition to the use of grammar and spelling. Furthermore, the questions with high scores can be isolated and from them and we can learn the common features among them by running a rule-based classifier.

7. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proc. of the 2008 International Conference on Web Search and Data Mining*, pages 183–194, 2008.

[2] D. Correa and A. Sureka. Chaff from the wheat: Characterization and modeling of deleted questions on stack overflow. In *Proc. of the 23rd International Conference on World Wide Web*, pages 631–642, 2014.

[3] J. Higgins. *The Radical Statistician: A Beginners Guide to Unleashing the Power of Applied Statistics in The Real World*. Jim Higgins Publishing, 2006.

[4] J. Jeon, W. B. Croft, J. H. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In *Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 228–235, 2006.

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

[6] Y. Yao, H. Tong, T. Xie, L. Akoglu, F. Xu, and J. Lu. Want a good answer? ask a good question first! *arXiv preprint arXiv*, 2013.

[7] A. Ying. Mining challenge 2015: Comparing and combining different information sources on the stack overflow data set. In *The 12th Working Conference on Mining Software Repositories*, 2015.