Predicting the quality of questions on Stackoverflow

Antoaneta Baltadzhieva
Tilburg University
a_baltadzhieva@yahoo.de

Grzegorz Chrupała
Tilburg University
g.a.chrupala@uvt.nl

Abstract

Community Question Answering websites (CQA) have a growing popularity as a way of providing and searching of information. CQA attract users as they provide a direct and rapid way to find the desired information. As recognizing good questions can improve the CQA services and the user's experience, the current study focuses on question quality instead. Specifically, we predict question quality and investigate the features which influence it. The influence of the question tags, length of the question title and body, presence of a code snippet, the user reputation and terms used to formulate the question are tested. For each set of dependent variables, Ridge regression models are estimated. The results indicate that the inclusion of terms in the models improves their predictive power. Additionally, we investigate which lexical terms determine high and low quality questions. The terms with the highest and lowest coefficients are semantically analyzed. The analysis shows that terms predicting high quality are terms expressing, among others, excitement, negative experience or terms regarding exceptions. Terms predicting low quality questions are terms containing spelling errors or indicating off-topic questions and interjections.

1 Introduction

CQA websites provide an interface for users to exchange and share knowledge. The user asking a question lacks knowledge of a specific topic and searches for an expert to provide the desired knowledge. In this way, the asker is querying a topic and the experts are the source of information, replacing other sources like documents or databases. However, the search results may not provide an exact solution to the user's problem. Although the idea of receiving a direct response to an information need sounds very appealing, CQA websites also involve risk as the quality of the provided information is not guaranteed. An important difference between user-generated content and traditional content is the range of the content quality: user-generated content shows a higher variance in quality (Agichtein et al., 2008) than traditional content (Anderson, 2006).

Stack Overflow (SO) is a CQA website in the field of computer programming. Access is free and answers are voted according to the asker's satisfaction[1]. The asker can tag a question to indicate a specific subject. Users can vote questions, answers and edits to indicate how helpful they were. The votes determine the user's reputation. In order to create a high-quality library of questions and their answers, SO allows users not only to post questions or answers but also to edit them.

Despite the encouragement of SO and the offered opportunities to maintain the content quality, a lot of questions on SO are not answered. With the increase in popularity of SO, not only the number of questions and the number of new members increased, but also the number of unanswered questions. According to statistics from 2012, approximately 45 questions per month remained unanswered (Asaduzzaman et al., 2013). By March 20, 2014, the number of unanswered questions was 752,533 out of 6,912,743 (approximately 10.9%). Interestingly, the fact that those questions are not answered is not caused by users not having seen them. In fact, unanswered questions are seen 139 times on average (Asaduzzaman et al., 2013). It is not obvious why a certain question receives more answers than others. Also, it is not clear whether the question characteristics

[1] http://stackoverflow.com/

that determine the number of answers a question receives also influence the question score. In this paper, we evaluate the features of questions in SO, how they influence the two above mentioned indicators of question quality, and attempt to predict these outcome measures for newly posted questions. Our main contributions are twofold. First, unlike previous work, we study the influence of specific individual terms, i.e. the words used to construct the question title and body. More specifically, we analyze the terms used in the posted questions and explore to what extent they can predict the question score and the probability of receiving an answer. The results indicate that the models have the best predictive power when the terms are included. Second, we study their influence on **two** measures of question quality: *the number of answers* and *the question score*.

1.1 Reserach overview

In the current study, we investigate which features influence question quality, as measured by the number of answers and the question score a question receives, in a programming CQA. Also, we predict which lexical terms determine high and low quality questions. We test the influence of question tags, length of the question title and body, presence of a code snippet and the user reputation on question quality. In addition, we test the influence of terms used to formulate the question. For each of the two dependent variables, we estimate Ridge regression models with an increasing number of independent variables on a dataset of over 1.7 million questions posted on Stack Overflow, dividing them into a training, validation and test set. The results indicate that the inclusion of terms in the models improves their predictive power. To the best of my knowledge, this research is the first to analyze the terms used in the posted questions and to explore to what extent they can predict the probability of receiving an answer. We rank the significant terms based on their coefficient value. The terms with the highest and lowest coefficients were semantically analyzed and divided in subgroups to gain a better understanding of the semantic nature of the terms. We find that terms predicting high quality are terms expressing excitement, negative experience or frustration, and terms regarding exceptions, or indicate that the questions are posted by new members. The largest groups of terms predicting low quality questions is the group containing spelling errors. Also words that mark off-topic questions and interjections are an indication of low quality questions. The better understanding of the terms used in low and high quality questions would help to improve the question formulation and herewith the content if CQA websites.

2 Related work

2.1 Question quality

Due to the large number of CQA websites, the importance of high-quality content in CQA websites has been recognized and investigated in several studies. Agichtein et al. (2008) found that there is a correlation between the question quality and answer quality, i.e. question quality will influence CQA service quality. According to Li et al. (2012), high quality questions are expected to draw greater user attention and will make users feel more compelling to answer the question within a shorter period of time.

Different studies employ different definitions of question quality. As measures of question quality we consider *the number of answers* and *the question score* as those are the response of the community to the usefulness of the question (Anderson et al., 2012). The number of answers is a direct feedback on the usefulness of the question. Research has shown it is the most significant feature to predict the long term value of a question together with its answers set (Anderson et al., 2012). Also the question score reflects the question quality. A question can be voted up or down by using the up or respectively down arrow on the left side of the question. In general, answered questions on SO have higher scores compared to unanswered questions (Saha et al., 2013).

Although the question score and the number of answers are considered quality determinants, they are not necessarily correlated. A question that addresses a new development which is interesting to the community but difficult to answer may receive no answers but a lot of upvotes. If however a question was too easy or posted previously it may receive answers, but may not be evaluated high as it does not contribute to the CQA. A number of other measures of question quality have been used in the literature. For a detailed overview of the existing literature, see (Baltadzhieva and Chrupala, 2015).

2.2 Features determining question quality

The features determining question quality are divided in a question-related and an asker-related group. The former is represented by the features *tags, terms, question title and question body length*, and *the presence of a code snippet*. Regarding asker-related features, the reputation of the user is taken into consideration. This researches focus on features that are available at the moment a question is posted, because features which are not available at the moment of the posting cannot help the asker to improve her question (Cheng et al., 2013; Correa and Sureka, 2014).

2.2.1 Question related features

In SO, askers can add tags to a question to indicate which topic(s) they address. Saha *et al.* (2013) analyzed the tags as topics and concluded that the large number of unanswered questions cannot be explained by a lack of sufficient experts for certain topics. Furthermore, Correa and Sureka (2014) observed that a high percentage of author-deleted questions are marked as too localized and off-topic, and that a high percentage of moderator-deleted questions are marked as subjective and not a real question. Asaduzzaman *et al.* (2013), state that incorrect tagging is one of the characteristics of unanswered questions. These results indicate that question topics, i.e. tags, may either be incorrect and/or may not be fully informative of the likelihood of receiving an answer, the number of answers, or question score. Therefore, a number of recent studies tried to infer question topics from the natural language used to formulate the questions. The current study uses both *tags*, as well as information from the questions' natural language formulation, the *terms*. In the term extraction process, terms are analyzed as the number of occurrences in the question title or question body where a term receives a value of 0 if it does not occur and otherwise the value of the number of occurrences.

Yang *et al.* (2011) found that the shortest and longest questions have the highest probability of obtaining an answer - short questions can be read and answered in a very short time, and long questions are mostly expertise-related, need more explanation and are therefore appealing for users with the same interest. In contrast, Asaduzzaman *et al.* (2013) found that too short questions are very likely to remain unanswered as they may miss important information; and too time-consuming questions are not very attractive for answerers. According to Saha *et al.* (2013) both classes have the same probability of receiving an answer. Correa and Sureka (2014), finally, found that compared to closed questions, deleted questions had a slightly higher number of characters in the question body. The existing literature is thus inconsistent regarding whether and to what extent question length influences question quality. Further, question length and question body length are never analyzed separately. Therefore, we explore the effects of both *question title* and *question body length* to see if the results point in the same direction.

Several studies have found that question categories that contain a code snippet have a high answer ratio and may have more than one possible good answer (Treude et al., 2011; Asaduzzaman et al., 2013). Also deleted questions have a lower percentage of code blocks than closed questions (Correa and Sureka, 2014). However, the presence of a code snippet may also have adverse effects as well if the code is hard to follow or if other users cannot see the problem (Asaduzzaman et al., 2013). Hence, it is unclear what the effect of *the presence of a code snippet* is on question quality.

2.2.2 Asker-related features

Regarding asker-related features, we consider the asker's reputation as a feature that influences question quality metrics. The reputation scores are built on users' participation on the CQA website. Users with high reputations do not only provide an essential contribution to CQA websites in general, but they also provide the most helpful answers (Welser et al., 2007; Pal et al., 2012). SO rewards upvotes on answers more than on questions and assigns high reputation users more privileges in site management and bonuses than regular users. The most reputation points are scored when a user's answer is accepted as the best answer, when it is upvoted or when the answer has received a bounty. Anderson *et al.* (2012) show that users build their reputation mainly by receiving upvotes for their answers and not by asking questions themselves. Saha *et al.* (2013) found the asker's reputation to be one of the most dominant attributes to distinguish between answered and unanswered questions, the former having a max score of twice as much as unanswered questions. For a detailed overview, see (Baltadzhieva and Chrupala, 2015).

3 Dataset description

Our dataset consists of JSON files extracted from SO using the Stack Exchange API (Application Programming Interface). The dataset contains questions in the period between 31 July 2008 and 9 June 2011. Within this time period, 1,713,400 questions were posted. Out of the total number of questions, 126,227 remained unanswered (7.37%). Each question contains information about the question itself, such as title, body, upvotes, downvotes etc., and about the question owner, e.g. registration status, reputation, name, id etc. In this research we are only interested in the variables as described below.

3.1 Data overview

Tables 1 and 2 provide an overview of the data and descriptive statistics of the key variables normalized.

Data item	Count
Questions total	1,713,400
Questions unanswered	126,227
Code snippet 1/0	792,822/920,578
Terms	36,865
Tags	12 11,613

Table 1: Data overview

	Mean	SD	Median
Nr. of answers	2.242	1.869	2
Q. score	1.331	2.446	1.00
Q. title length	8.27	3.71	8.00
Q. body length	91.74	87.43	72.00
User rep.	1600.40	5552.40	301.00

Table 2: Descriptive statistics

The independent variables *title length, body length* and *user reputation* are normalized by the logarithmic transformation using the natural logarithm, and for *question score* and *number of answers* we use percentile normalization. Most questions receive a small number of answers. On average a question receives a relatively low score. Question titles and bodies consisting of only one word may be questions where only a code snippet was posted. The high mean value of the user reputation suggests that many SO users have a high user reputation. As it has been shown that there is a positive relationship between the user reputation and how fast the user replies to a question (Anderson et al., 2012), it can be concluded that SO askers are active community users.

To predict the number of answers and question score, the independent variables are defined as follows: the tags and the presence of a code snippet are represented as a Boolean value; the question title and body length are measured by the number of words; the user reputation is the user reputation score; the terms are a count variable of how often a term occurs in the question title or body. In order to extract numerical information from text content, first a tokenization process takes place (Manning et al., 2008). Stop words are filtered out from the vocabulary prior to natural language processing, because they are of little value in finding documents matching a user's information need (Manning et al., 2008).

Only the tags are included that appear in at least 20 questions and terms that appear at least in 50 questions. Results based on tags and terms that occur seldom are likely to be spurious and are not expected to have strong predictive power.

3.2 Method of analysis

For the prediction task we use multiple linear regression models. The expected relationship is a linear function of the independent variables (Field, 2009):

$$y_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} + \epsilon_i$$

Here, for question i, y_i represents the dependent variable *question score* or *number of answers* received, β represents the coefficients of the predictor variables x and ϵ is the difference between the predicted and the observed value of the outcome variable, which is assumed normally distributed.

When predicting future responses and investigating the relationship between the response variable and the predictor variables regularized regression models are preferred, because they solve highly variable estimates of the regression coefficients when there is multicollinearity or when the number of predictors is very large in connection to the number of observations (Hartmann et al., 2009). In programming languages a lot of terms appear together what can lead to multicollinearity. As the number of terms used in this study is extremely large (36,865) and in order to avoid overfitting, a regularized regression model, Ridge regression (Hoerl and Kennard, 1970b; Hoerl and

Kennard, 1970a), is used. Ridge regression applies a penalty to the sum of the squared values of the regression coefficients which shrink the coefficients towards zero, but never become zero, which means that all predictors remain in the model. Applying a penalty results in lower expected prediction error because it reduces the estimation variance (Hartmann et al., 2009).

We split the dataset in three subsets: a training set - the first 60%, a validation set - the next 20%, and a test set - the rest 20%. The sets are chronologically partioned as the goal of this study is to predict the quality of new questions. The validation set is used to optimize the regularization parameter for each model. To find the optimal ridge parameters, several values are tried in increasing order. The value that reduces the Mean Squared Error of the validation set the most is chosen as the optimal parameter. Finally, the obtained coefficients, given the optimal regularization parameter, are applied on the test set to assess the predictive validity of the models.

To investigate the question quality, two sets of multiple linear regression models are applied – one to predict the question score and the second one to predict the number of answers. For each set, four different regression models are applied and compared in order to discover which independent variables have the most predictive power. Model 0 is the baseline intercept-only model. In Model 1 only the tags are included, Model 2 contains question title and body length, code snippet and tags, Model 3 - the variables of Model 2 plus user reputation, and Model 4 - the variables of Model 3 plus terms. Each set uses the same dependent variable and a different set of independent variables. To compare the performance of the models in each set, the R-squared, Mean Squared Error (MSE) and Mean Absolute Error (MAE) are reported.

3.3 Results

As Model 4 has the best performance on the *test set* as presented in Table 3 and Table 4, only the coefficients of Model 4 are discussed in this section.

The results show that Model 1 performs better than the baseline model for both question score and number of answers, as the MSE has lower values. Compared to Model 2 however the performance does not change drastically. The MSE of Model 2 for predicting question score decreases with only 0.075 and for number of answers with only 0.031. These results indicate that the tags do influence the question quality, whereas the inclusion of the title length, body length and the presence of a code snippet gives a minor improvement.

	MSE	**MAE**	$\mathbf{R^2}$	**F-statistic**
Model 0	5.675	1.482		
Model 1	5.138	1.375	0.088	3.768
Model 2	5.063	1.363	0.102	4.396
Model 3	4.869	1.323	0.136	6.124
Model 4	4.622	1.286	0.180	1.897

Table 3: Ridge regression *question score*

	MSE	**MAE**	$\mathbf{R^2}$	**F-statistic**
Model 0	3.199	1.353		
Model 1	2.769	1.228	0.109	4.771
Model 2	2.738	1.219	0.119	5.257
Model 3	2.630	1.192	0.154	7.079
Model 4	2.514	1.163	0.191	2.048

Table 4: Ridge regression *number of answers*

For both model sets it applies that the more complex the model is, the better it performs on the training and test set: MSE and MAE decrease with the increase of the number of independent variables and all models outperform the baseline Model 0. This implies that Model 4 for both question score and number of answers fits the data best. The same conclusion can be drawn from the R-squared values. For number of answers, the R-squared for the test set increases from 0.102 for Model 2 to 0.180 for Model 4, meaning that Model 2 explains 10.2% of the variance in the question score in the test set while Model 4 explains 18.0%. Similarly, with regard to number of answers, the R-squared values for Models 1 and 3 for the test set are 0.119 and 0.191, respectively.

3.4 Coefficient analysis

As Model 4 has the best performance, only the coefficients of Model 4 are presented and discussed. The *question title and body length* and *the presence of a code snippet* have a significant negative effect on the outcome variables, while reputation has a positive effect. To better understand the effect size, we calculate the effect of a 10% increase in body length, title length and user reputation, while taking the natural logarithm into account. A 10% increase in *title length, body length* and

user reputation results in a change in the questions score of -0.010, -0.019 and 0.015, respectively. Including a code snippet reduces the question score by -0.155. Hence, the effect of all variables is fairly small. In Model 4, for *number of answers*, the *title length* effect is $\beta_t l$ = -0.058, which implies that, taking the mean title length as baseline and accounting for the logarithmic transformation, a 10% increase in *title length* results in a 0.006 reduction in the number of answers. Similarly, a 10% increase in *body length*, $\beta_b l$ = -0.132, and *user reputation*, $\beta_u r$ = 0.122, gives an increase in the number of answers of -0.013 and 0.012 respectively. Including a code snippet reduces the expected number of answers by -0.050. The effects of the predictors are again fairly small.

3.4.1 Parts of speech

Excessive use of (only) one part of speech might also influence the question quality. For example, too many verbs in a sentence can make it sound heavy and wordy (Weber, 2007) and therefore unpleasant to read. The number of nouns, verbs and adjectives are calculated using the Natural Language Toolkit (NLTK)[2]. Most of the terms that predict question score are nouns - 53.55%. This is not surprising as nouns are used most frequently in natural language. For number of answers, a Chi-square test is used to show that the counts of parts of speech differ significantly between high and low quality questions ($\chi^2 = 37.362$, $df = 3$, $p = 0.01$). Particularly, the percentage of nouns is higher in the groups of terms predicting low question quality – 65.04%. At the same time the percentage of used adjectives is higher for high question quality – 13.55% vs. 8.98% for low questions quality. As adjectives are words that have a descriptive character and are used to assign a noun a specific property, it may be concluded, that questions with a low number of answers are less descriptive and maybe do not explain the information need clearly enough. For question score, the counts of parts of speech do not significantly differ between the high and low quality groups ($\chi^2 = 1.190$, $df = 3$, $p = 0.755$).

3.4.2 Semantic analysis

In the term analysis only terms were included that have a statistically significant influence on the question quality. Due to the large number of such terms, we analyze only 10% of the terms with the highest coefficient values as they contribute to high question score and number of answers and 10% of the terms with lowest coefficient values that determine questions with low score and low number of answers. We assume that this percentage provides enough terms to discover patterns.

The extracted terms are analyzed and first divided into two groups – professional/expertise terms and generic terms. We assume that the question subject is expressed by the tags and that professional/expertise terms would overlap often with the tags. Furthermore, the goal of the study is not to explore the question topics, but the lexical terms. Therefore, only the generic terms will be considered and subdivided into several semantic groups. To be able to make a distinction between the two groups, in the programming/expertise term set, we include strict programming/expertise terms such as *resig, dataframe*, and words that are considered expertise words, not commonly used in natural language conversation such as *deprecate, indention* etc. We use the SO website for additional reference to recognize expertise terms, such as *mythical* that refers to the Software Engineering book The Mythical Man-Month by Fred Brooks (1975) or *girlfriend* that refers to the programming website Cocoa is my Girlfriend[3]. As proper nouns are mostly used as a reference and link to a new information source, they are considered too general and added to the group of generic terms.

The analysis shows that, for both high and low quality questions, the generic terms dominate. The terms having the most predictive power for *number of answers* are: *pricey, tolerable, fascinated, aspiring, believer, addicted, contenders, advocates, argues, laughing, praise, religious, corey, sniffed, motivations, analogies, techie, geeky, internationally, misconceptions*. The twenty most predictive terms for question score are: *fascinated, addicted, praise, mentality, camps, rage, lippert, misconceptions, blatant, contenders, mandated, analogies, coolest, speculate, thoughtful, newcomers, picturing, stackers, replays, darned*. For both dependent variables, we test whether there is a significant difference in the counts of generic and professional/expertise terms between high and low quality questions. Chi-square tests indicate that the differences are significant: χ^2 = 6.833, df = 1, $p < 0.01$ for question score and χ^2 = 24.189, df = 1, $p < 0.01$ for number of answers. For both de-

[2] www.nltk.org

[3] http://www.cimgf.com/

pendent variables we see the same pattern: in the term group that contributes to low question quality, the number of programming/expertise terms is larger. To have a better understanding of the nature of the generic terms, a further distinction was made based on the semantic nature of the terms.

The terms predicting a high question quality, can be divided in subgroups where the following subgroups are very similar across the two dependent variables:

Category	Examples
Excitement	praise, compelling, thrilled
Neg. Experience	blatant, miserable, horrific
Discussion	speculate, agree, misguided

Table 5: Semantic categories

The group of Excitement consists of terms which describe a passionate attitude towards a programming problem. These terms are assumed to be used by users who express emotional commitment to the subject in question. Terms of excitement that predict high question score are *fascinated, compelling, praise, remarkably, aspiring* etc. Similarly, terms such as *thrilled, believing, passion, amazed, enjoyed* account for a higher number of answers. The group of Negative experience/Frustration group consists of terms which express a negative emotion, mostly caused by lack of success when trying to solve a specific problem, i.e. *blatant, miserable, darned, disastrous, insanity, dread etc.* which, according to the model results, indicate high question score. Examples of terms of negative experience or frustration that account for high number of answers are *horrific, miserable, torn, scare, evil* etc. Such high degree of frustration may be the results of multiple attempts to solve the problem which indicates that the user is providing a serious question. The third group lists terms that are used to start a discussions or explanations of a particular problem: *speculate, agree, disagree, advocate, argumentative* suggest an attempt to discussion, and *beware, misguided, unambiguous* assume that a user is trying to explain a specific issue. Although the words in this group seem related, they are less distinct and further research should perform a more in-depth analysis of this group.

We found two more subgroups that account for a high question score:

The former determines questions posted by new members. Apparently, when users admit that they are new in the programming world, their question is appreciated by other new users or welcomed by experienced users who remember their first programming steps; or they are just easy to answer. The terms in the Exceptions group are used to discuss exceptional programming issues - *peculiarity, obscurity, surprises, counterintuitive, unintentional, nontrivial, contradicting, unintuitive*. Such cases seem to be intriguing and challenging for the community and are therefore more likely to be appreciated and highly graded.

Category	Examples
New members	newbies, newcomers, freshman
Exceptions	peculiarity, obscurity, surprises

Table 6: Semantic categories

The following categories have negative effect on the question quality:

Category	Examples
Spelling errors	workin, acessing, specifc
interjections	hmmm, hay, aha
Off-topic terms	hiring, graduate, bosses

Table 7: Semantic categories

The terms that have a negative effect on the question score and the number of answers have one subgroup in common - the group of the misspelled words. In the group of terms predicting a low number of answers 8.31% is not spelled correctly. It can be assumed that questions containing typos are not considered professional and worthy for the community. Such questions may not be taken seriously and users may refuse to spend time giving an answer. More importantly, terms containing typos would not appear in the search results. Apparently, SO users often ignore the integrated spelling checker. In the group of terms having a negative effect on the number of answers, also off-topic terms and interjections that express sounds normally used in daily conversations and more common in speaking than in writing were found. To the off-topic group belong terms that are used mostly in questions related to people searching for or offering a job, students searching for answers to problems for their bachelor thesis. Such questions may be considered as off-topic and not worthy to community users.

4 Discussion

The aim of this study is to investigate to what extent the discussed features influence the number of answers and the question score a question receives, and whether it is possible to predict these measures of question quality. The results from both sets of models showed that the inclusion of linguistic information improves the prediction accuracy of the models. An analysis of the extracted terms shows that they can be classified in subgroups based on their semantic nature. First, certain groups of generic terms have greater impact on question quality. Second, questions that contain terms regarding newcomers, attempts at discussion or explanation of a problem or strong commitment to the problem are more likely to receive a high question score and a large number of answers. Finally, the questions that are considered not worthy of a positive evaluation or receiving an answer are questions that include typos or that are found to be off-topic.

These findings are in line with Correa and Sureka (2014) and Saha *et al.* (2013) who find that deleted questions in SO are questions that are considered poor quality and off-topic. Also Saha *et al.* (2013) found that homework and job-hunting belong to the tags in deleted questions.

Another clear characteristic of low quality questions are misspellings and typos. Online social media sources are often characterized by not following common writing rules (Agichtein et al., 2008). Not taking them into account seems not appreciated and considered unprofessional.

With regard to the terms predicting high quality questions, the results of the current research revealed more similarities. Nasehi *et al.* (2012) considered the following question types groups: debug/corrective, need to know, how-to-do-it, seeking different solution. Truede *et al.* (2011) distinguish similar groups – decision help, error, how-to, discrepancy, review. All of these questions can be seen as seeking an explanation. To present their information need, askers use terms like *speculate, agree, disagree, argues* which were found to have a significant positive effect on the question quality.

Existing literature does not provide a consistent explanation of whether a code snippet increases the question score or the number of answers. Our study showed that the effect of a code snippet is negative which is in line with the statement of Asaduzzaman *et al.* (2013) who explained that a code snippet may have a negative effect on the number of answers if the code is hard to follow or the problem is not clear.

There also is disagreement in previous work about the influence of the question title and question body length. Where some researchers stated that very short and very long question are more likely to obtain an answer (Yang et al., 2011), others found that too short questions may miss important information and may therefore remain unanswered (Asaduzzaman et al., 2013). Our study indicates that the length variables negatively affect question quality. The current results thus are mostly in line with the findings of Correa and Sureka (2014) who found that deleted questions have a higher number of characters in the question body than closed questions. Although, title length, body length and the inclusion of a code snippet all have significant negative effects on the question quality, it must be noted, that all effects are rather small.

Regarding the quality measure user reputation, our results are in line with previous work. As Yang *et al.* (2011) also showed, users with a high reputation are more likely to receive an answer than new users who logically have a lower reputation. For both, question score and number of answers, it was found that the higher the reputation, the higher the value of the quality measure.

5 Future research

In the current study lexical entities, the terms, are included to predict question quality above the level of the assigned tags. However, the terms were analyzed manually, based on human judgment. This is rather subjective and may result in a somewhat arbitrary assessment. An automated way to analyze the extracted terms would be an improvement and a good suggestion for future research. Another matter for a future work is to include the part-of-speech tagging in the predicting models and to use the parts of speech as features to improve the predictive power of the models.

References

Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 183–194. ACM.

Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg,

and Jure Leskovec. 2012. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 850–858. ACM.

Chris Anderson. 2006. *The long tail: Why the future of business is selling less of more*. Hyperion.

Muhammad Asaduzzaman, Ahmed Shah Mashiyat, Chanchal K Roy, and Kevin A Schneider. 2013. Answering questions about unanswered questions of stack overflow. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, pages 97–100. IEEE Press.

Antoaneta Baltadzhieva and Grzegorz Chrupala. 2015. Predicting question quality in question answering forums.

Frederick P Brooks. 1975. *The mythical man-month*, volume 1995. Addison-Wesley Reading, MA.

Derrick Cheng, Michael Schiff, and Wei Wu. 2013. Eliciting answers on stackoverflow.

Denzil Correa and Ashish Sureka. 2014. Chaff from the wheat: characterization and modeling of deleted questions on stack overflow. In *Proceedings of the 23rd international conference on World wide web*, pages 631–642. International World Wide Web Conferences Steering Committee.

Andy Field. 2009. *Discovering statistics using SPSS*. Sage publications.

Armin Hartmann, Anita J Van Der Kooij, and Almut Zeeck. 2009. Exploring nonlinear relations: models of clinical decision making by regression with optimal scaling. *Psychotherapy Research*, 19(4-5):482–492.

Arthur E Hoerl and Robert W Kennard. 1970a. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Arthur E Hoerl and Robert W Kennard. 1970b. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Baichuan Li, Tan Jin, Michael R Lyu, Irwin King, and Barley Mak. 2012. Analyzing and predicting question quality in community question answering services. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 775–782. ACM.

Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. 2008. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

Seyed Mehdi Nasehi, Jonathan Sillito, Frank Maurer, and Chris Burns. 2012. What makes a good code example?: A study of programming q&a in stackoverflow. In *Software Maintenance (ICSM), 2012 28th IEEE International Conference on*, pages 25–34. IEEE.

Aditya Pal, F Maxwell Harper, and Joseph A Konstan. 2012. Exploring question selection bias to identify experts and potential experts in community question answering. *ACM Transactions on Information Systems (TOIS)*, 30(2):10.

Ripon K Saha, Avigit K Saha, and Dewayne E Perry. 2013. Toward understanding the causes of unanswered questions in software information sites: a case study of stack overflow. In *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*, pages 663–666. ACM.

Christoph Treude, Ohad Barzilay, and Margaret-Anne Storey. 2011. How do programmers ask and answer questions on the web?: Nier track. In *Software Engineering (ICSE), 2011 33rd International Conference on*, pages 804–807. IEEE.

Wibke Weber. 2007. Text visualization-what colors tell about a text. In *Information Visualization, 2007. IV'07. 11th International Conference*, pages 354–362. IEEE.

Howard T Welser, Eric Gleave, Danyel Fisher, and Marc Smith. 2007. Visualizing the signatures of social roles in online discussion groups. *Journal of social structure*, 8(2):1–32.

Lichun Yang, Shenghua Bao, Qingliang Lin, Xian Wu, Dingyi Han, Zhong Su, and Yong Yu. 2011. Analyzing and predicting not-answered questions in community-based question answering services. In *AAAI*.