Mining Successful Answers in Stack Overflow

Fabio Calefato, Filippo Lanubile, Maria Concetta Marasciulo, Nicole Novielli
Dipartimento di Informatica, Università degli Studi di Bari - Bari, Italy
{fabio.calefato, filippo.lanubile, nicole.novielli}@uniba.it, {m.marasciulo2}@studenti.uniba.it

***Abstract*— Recent research has shown that drivers of success in online question answering encompass presentation quality as well as temporal and social aspects. Yet, we argue that also the emotional style of a technical contribution influences its perceived quality. In this paper, we investigate how Stack Overflow users can increase the chance of getting their answer accepted. We focus on actionable factors that can be acted upon by users when writing an answer and making comments. We found evidence that factors related to information presentation, time and affect all have an impact on the success of answers.**

***Index Terms* —** *Online Q&A, Sentiment Analysis, Knowledge Sharing, Human Factors.*

I. Introduction

The enormous success of Stack Overflow (SO) provides data scientists with a huge amount of data about online question answering (QA). Our investigation aims to provide guidelines for writing high-quality contributions and inform the design of tools that support effective knowledge sharing. In this paper, we investigate how an information provider can increase the chance of getting his answer accepted in SO. In particular, we focus on actionable factors that can be acted upon by community members when contributing to answering a question. Hence, our first research question is formulated as follows:

RQ1 – Which actionable factors predict the success of a SO answer?

Social and temporal aspects are among the success factors of an answer [1][4], depending on the answerers' level of expertise and their engagement in the community. More recently, research has begun to investigate linguistic factors too, looking at how answers are formulated [5][7]. In addition, we argue that the path to effective question answering and reputation building passes through emotions too. There is an increasing attention to the impact of emotional awareness on effective collaboration [5][8]. However, existing research on online QA sites has not taken into full consideration the potential contributions from the field of affective computing, with the only notable exception of a large-scale sentiment analysis study on Yahoo! Answers [9]. Therefore, we formulate our second research questions:

RQ2 – Do affective factors influence the success of a SO answer?

While previous research has mostly focused on time, reputation and presentation quality, our study is the first one to investigate the impact of affective factors on the success of answers in SO. This study is part of our ongoing research on investigating the role of emotions in community-based QA, and their impact on effective knowledge creation and sharing [11].

II. Success Factors for Answers

The actionable factors in our model include presentation quality, affect and time. Social factors are also added as a control dimension, due to the evidence of their impact on success [1][4]. In the following, we present the predictor variables for each factor in our framework.

A. Presentation Quality

Writing a good answer in SO involves successfully complying with the community standards of presentation quality. SO uses a set of simple textual metrics to pre-filter low quality posts including *Length* (# characters), *Uppercase Ratio*, and *URL Count.* Therefore, we included these metrics in our framework to capture the presentation quality of an answer. Previous research [15] also found that code snippets affect the success of SO questions. We then consider the *Presence of Code Snippets* (binary variable) as a predictor for the success of SO answers.

B. Affect

Displaying emotions is common in face-to-face interaction. However, people might not be prepared for effectively dealing with the barriers of social media to non-verbal communication. This clearly emerges in discussions where users complain about harsh comments from experts[1]. Moreover, SO guidelines include a 'Be nice' section[2] in which users are invited to be patient and avoid offensive behavior. Furthermore, previous research on success of questions has shown how strong negative emotions in follow-up discussions discourage participation [2]. Accordingly, we consider textual cues for affective states among the potential factors of success for an answer. Specifically, we consider metrics describing the overall polarity (positive vs. negative) and intensity of the sentiment expressed in an answer (*Answer Positive/Negative Sentiment*) and related comments until the acceptance (*Comment Positive/Negative Sentiment*).

To capture the sentiment of answers and comments, we use SentiStrength [14], a state-of-the-art tool already employed in social computing [6][9], which is capable of dealing with short written informal, including abbreviations, intensifiers and emoticons. Based on the assumption that a sentence can convey

[1] http://meta.stackexchange.com/questions/179003/stack-exchange-is-too-harsh-to-new-users-please-help-them-improve-low-quality-po/179009#179009
[2] http://stackoverflow.com/help/be-nice

mixed sentiment, SentiStrength outputs both positive and negative sentiment scores for an input text. It assigns the overall positive and negative scores to a text by considering the maximum among all the sentence scores**.** Positive sentiment scores range from +1 (neutral) to +5 (extremely positive) while negative sentiment scores range from −1 (neutral) to −5 (extremely negative). In our analysis, we adjust the sentiment score and map them into the [0,4] interval, with zero indicating the absence of positive or negative sentiment.

C. Time

Speed is among the key factors of success of SO. The median time for a first answer is only 11 minutes (21'20'' for an accepted answer) [10]. Expert users are also the fastest contributors [1], resulting in a high probability of askers accepting the first answer [4]. Moreover, it has been observed that the longer the wait to get the first answer the less likely is for an answer to be eventually accepted [1]. Therefore, we include the arrival order (*Ranking*) and the time, in seconds, from the moment the question is posted (*Elapsed Time*).

D. Reputation

The social reputation system of SO is designed to incentivize contributions and allow assessment of trustworthiness of users. The *Answerer's Reputation Score* and *Number of Badge*s earned is always publicly displayed along the post and then they may have an impact on the perception of quality of the answer. Moreover, previous research has shown how the users with a high reputation are more effective in providing successful answers [1][4]. Therefore, we include the social reputation of the answerer as a control factor in our model. Moreover, since new users may not be familiar with the community rules involving the acceptance of the best answer, we also add the *Asker's Reputation Score*.

III. Dataset

We extracted our dataset from the official SO data dump, updated on September 2014 [16]. SO official datasets always report the reputation score of users when the dump is created. The SO reputation system assigns users to the following categories, based on the reputation score gained: *New Users* (score <10), *Low Reputation Users* (score in [10,1000[), *Established Users* (score in [1000,20k[), and *Trusted Users* (score ≥ 20k). Since SO allows users to gain at most 200 reputation points per day, it is reasonable to assume that the reputation category of the largest majority of users stays unvaried over a month [4]. Therefore, we built the dataset for our analysis by considering the answers to the questions posted during the last month of the dump (14^{th} Aug.-14^{th} Sept.). To enlarge our dataset, we also consider answers to the questions posted from the 5^{th} of April to the 5^{th} of May, corresponding to the last month of the previous dump of May 2014. We do not consider answers to questions that are removed or closed by moderators. We then obtain a total of 439,586 answers, from which we also remove self-answers (32,687), mainly provided by new users[3]. We further remove answers that were edited after the acceptance vote (58,281) because we cannot know what an answer looked like at the acceptance time if it has been further modified. The final dataset resulting from pre-processing contains 348,618 answers, of which 103,728 (30%) accepted.

IV. Results

We model the success probability of an answer in a logistic regression framework since it allows us to reason about the significance of one factor given all the others. We use the acceptance vote as the dependent variable and presentation quality, affective, temporal, and social metrics as independent variables.

Results (Table I) are obtained by randomly splitting the dataset into training (70%) and test (30%) sets, while keeping the same percentage of successful questions as in the entire dataset. We run the classification experiment using Weka[4]. We assess the prediction quality in term of Receiver Operating characteristic Area Under Curve (AUC), which is the recommended approach for evaluating binary decision problems [12]. As a baseline we choose AUC = 0.50, corresponding to the curve associated to random prediction. We perform our classification experiment in an ablation test setting by removing one of the success factors at a time, while retaining the others. Thus, through the decrease of classification performance, we evaluate the contribution of each factor to the prediction task. For each setting, we report the percentage of decrease of AUC with respect to the complete feature setting (All). To test the statistical significance of differences of each model with respect to the complete one, we perform the Delong's test for correlated ROC curves using the pROC package for R [13]. For the sake of completeness, we also provide weighted F-measure, Precision and Recall.

Although our best model is far from perfect (0.6498) all models significantly improve upon the random baseline ($p<0.05$). As expected, *Time* is the most predictive factor causing the highest drop in recognition performance, followed by *Reputation* and *Presentation Quality*.

TABLE I. Prediction Results for Logistic Regression Models Using Different Sets of Features.

Setting	AUC	%Decr AUC	F	Prec	Rec
All	.6498		.6043	.6479	.7064
-w/o Reputation	.6371	-1,96% *	.5874	.6216	.7061
-w/o P. Quality	.6389	-1,69% *	.6019	.6492	.7067
-w/o Time	.5912	-9,03% *	.5984	.6394	.7056
-w/o Affect	.6500	0,03%	.6045	.6501	.7067
* AUC curve significantly different from the complete setting (All), with $\alpha = 0.05$					

[3] The 90% of accepted answers provided by new users are self-answers.

[4] http://www.cs.waikato.ac.nz/ml/weka/

To exhaustively investigate the role that *Presentation Quality* and *Affect* play, we repeat the logistic regression analysis on a reduced dataset in which we include only the answers provided until the accepted one has been posted. This filtering results in a new dataset of 311,733 answers of which 103,728 (33%) accepted. The results of logistic regression on this dataset are reported in Table II. We use a standard Likelihood Ratio test to compute significance scores.

Reputation. Consistently with literature [1][4], a positive association is observed between the reputation score of an answerer's and the probability of success. The asker's reputation score also influences positively the chance of getting an answer accepted. In search for an explanation, we studied the distribution of the badge 'Scholar' among all users who have asked at least a question. Users unlock this badge when accepting an answer for the first time. We found that only the 29% of New Users holds the scholar badge. The percentage increases with the reputation score (up to 72%, 97% and 99% for Low Reputation, Established and Trusted Users, respectively) thus confirming the tendency of less expert users to not accept answers albeit useful.

Presentation quality. Including a code snippet greatly increases the probability of success (largest parameter estimate). The same finding applies to providing contextual information using URLs, although at a lower extent. Length has a weak impact on the probability of success, thus confirming evidence provided by previous work [7].

TABLE II. RESULTS OF LOGISTIC REGRESSION FOR PRESENTATION QUALITY, AFFECT AND SOCIAL FACTORS

VARIABLE	COEFFICIENT ESTIMATE	ODDS RATIO	p < 0.05
Intercept	-2.17	0.1142	*
Presentation Quality			
Number of URLs	0.0760	1.0790	*
Presence of Code Snippet	0.2800	1.3231	*
Length	0.0004	1.0004	*
Uppercase Ratio	-0.2933	0.7458	
Affect			
Answer Positive Sentiment	0.0214	1.0216	
Answer Negative Sentiment	0.0434	1.0444	*
Comment Positive Sentiment	0.0906	1.0948	*
Comment Negative Sentiment	-0.0377	0.9630	*
Reputation			
Answerer's Reputation Score	2.19E-06	1.0000	*
Answerer's Number of Badges	0.0002	1.0002	*
Asker's Reputation Score	4.74E-06	1.0000	*

Affect. Out of all the four variables considered in the affective dimension, we find that the positive sentiment expressed in the answerer's comments has the most significant impact on success, immediately followed by the negative sentiment within comments that, conversely, negatively influences the probability of getting an answer accepted. Among the 311,733 answers in our dataset, only 64,043 (21%) originate a follow-up discussion involving also the answer's author. Still, the impact of sentiment expressed in comments is significant.

To further investigate this phenomenon, we analyze the top 100 answerers' comments with the highest positive and negative scores, respectively. This analysis reveals that comments are very rich in affective lexicon. We observe that SO users express a wide variety of affective states in comments. We speculate that this might occur because comments are seen as a 'free zone' since reputation mechanisms do not apply to comments. More in detail, as for positive comments we found that the main affective states are gratitude (e.g., '*Thanks for the feedback, it was a pleasure!*'), wishes (e.g., '*Happy coding!*') and positive feelings linked to satisfaction and happiness for the help provided (e.g., Asker: 'Thanks, that helped. Case closed!' – Answerer: '*Thanks for the feedback! It was a pleasure!*'). As for negative comments, we observe a wider variety of emotions, ranging from negative attitude towards the asker to negative sentiment involving a positive attitude towards the reader (see Table III) A typical example of an offending comment is reported in the following discussion:

Asker: '*Worked perfectly, thanks a lot… one question if I want to do it using do end block, how can we do that?*' Answerer: '*Looks like you have changed your needs. That's disgusting*' – Asker: '*Fundamental are same, I just changed the format after you answered the first query. I don't see why this is disgusting. Anyways your solution was helpful… I read your blog after that I was not expecting such comments from you…*'.

We observe that the asker confirms that the answer was helpful. However, the asker is clearly disappointed by the negative attitude of the answerer, as reported in the comment. In fact, the answer is not finally accepted.

Negative sentiment is also detected when people apologize for not being able to provide further help (see positive attitude examples in Table III). Furthermore, users employ a negative lexicon also for expressing opinions on controversial technical issues (see Table III). These findings suggest that sentiment analysis could be exploited in detecting questions that have not been exhaustively answered, which is a goal addressed by research on effective knowledge-sharing in SO [1].

As for the emotional style of answers, negative sentiment is positively associated to success, which is unexpected and in contrast with what observed for comments. In order to provide an explanation, we analyze the top 100 answers with the highest negative sentiment score. We observe that a negative sentiment is actually expressed in the great majority of these answers. However, the negative polarity of these answers does not involve a negative judgment of the asker but it rather results in either an attempt of showing empathy towards the asker (e.g., '*If you are really worried about storage usage […]*', '*This could be very annoying but it is simply solved*', '*[…] This will make your experience a lot less frustrating*') or in a criticism towards a technological issue (e.g., '*This could work, but feels really awful*', '*This is extremely ugly for loop construction*').

Finally, due to the presence of domain-specific lexicon, examples of false positives in negative sentiment detection

emerge from the analysis of both comments and answers, such as in '*You are vulnerable to this bug*' or '*Kill this process*'. The domain-dependency of sentiment analysis tools is a known problem [3], meaning that applying a tool outside the domain in which it was tested, may produce unreliable results.

TABLE III. EXCERPTS OF ANSWERERS' NEGATIVE COMMENTS

NEGATIVE POLARITY, NEGATIVE ATTITUDE
'Didn't notice the ***horrid*** inline jQuery' 'Added some instructions for the ***really hopeless*** cases' '***Arrrghhh***, how I ***hate*** those people who downvote answers without leaving a comment as for why the downvote…
NEGATIVE POLARITY, POSITIVE ATTITUDE
'To explain my ***regrettably unfriendly*** comment (***sorry*** about that) […]' 'You could try this one (not optimal, I am ***afraid***)' '*I'm* ***afraid*** *I can't help you any further with this issue!'*
OPINIONS ABOUT CONTROVERSIAL TECHNICAL ISSUES
Asker: 'But what if you do have to worry about spaces in your filenames?' – Answerer: 'Then ***you've got major problems***! Let me meditate on it; ***it is extremely unpleasant***, whatever. [..]'
Answerer: 'Sorry for all the editing but this is a ***ridiculously complex*** issue'.

V. CONCLUSIONS

We performed an empirical investigation on the impact of presentation quality, affective, temporal, and social factors on the success of SO answers. Our study provides evidence-based guidelines that users can follow to increase the chance of getting their answers accepted. This study is the first one to address how the affective twists expressed in the language may influence the success of a contribution.

Our results confirm previous findings about the importance of being prompt in replying. Good presentation quality also fosters success, with presence of code snippet resulting as the best predictor of an answer acceptance.

Furthermore, we demonstrate the importance of considering affective factors among the predictors of success. Based on our findings, we conclude that providing help is a two-phase activity, where writing an answer is just the first phase. We observe how it is desirable for a contributor to be nice to the asker, especially in the follow-up discussions originated by his post. We believe that avoiding a negative attitude in comments could improve not only the contributor's chance of getting his answer accepted but also it could foster the involvement in the community of new users, which are often discouraged to contribute by the unsympathetic comments of experts.

While being preliminary, the findings of our study suggest directions for future work. First, the wide variety of affective states expressed in comments suggest the need for further investigation of the role of emotions in SO by considering more fine-grained emotion analysis since different emotions might be relevant to different contexts and tasks. Moreover, we underline the need for tuning state-of-the-art resources for sentiment detection by adapting them to domain-dependent use of lexicon.

ACKNOWLEDGMENT

The computational work has been executed on the IT resources of the ReCaS and PRISMA projects, financed by MIUR.

REFERENCES

[1] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. 2012. Discovering value from community activity on focused question answering sites: a case study of stack overflow. Proc. of 18th ACM SIGKDD Internat'l Conference on Knowledge Discovery and Data mining (KDD '12). ACM, 850-858.

[2] M. Asaduzzaman, A.S. Mashiyat, C.K. Roy, K.A. Schneider. 2013. Answering questions about unanswered questions of Stack Overflow, Proc. of the 10th IEEE Working Conf. on Mining Software Repositories (MSR 2013), 97-100.

[3] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. 2005. Pulse: Mining customer opinions from free text. Lecture Notes in Computer Science, 3646, 121-132.

[4] A. Bosu, C. S. Corley, D. Heaton, D. Chatterji, J. C. Carver, and N.A. Kraft. 2013. Building reputation in StackOverflow: an empirical investigation. Proc. of the 10th Working Conference on Mining Software Repositories (MSR '13). IEEE Press, 89-92.

[5] G. Gkotsis, K. Stepanyan, C. Pedrinaci, J, Domingue, and M. Liakata. 2014. It's all in the content: state of the art best answer prediction based on discretisation of shallow linguistic features. Proc. of the 2014 ACM conference on Web science (WebSci '14). ACM, 202-210.

[6] E. Guzman, D. Azócar, and Y. Li. 2014. Sentiment analysis of commit comments in GitHub: an empirical study. Proc. of the 11th Working Conference on Mining Software Repositories (MSR 2014). ACM, 352-355.

[7] K. Hart and A. Sarma. 2014. Perceptions of answer quality in an online technical question and answer forum. Proc. of the 7th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE 2014). ACM, 103-106.

[8] A. Kolakowska, A. Landowska, M. Szwoch, W. Szwoch, M.R. Wrobel. 2013. Emotion recognition and its application in software engineering, Proc. of the 6th Internat'l Conference on Human System Interaction (HSI), 2013, 6-8, 532,539.

[9] O. Kucuktunc, B.B. Cambazoglu, I. Weber, and H. Ferhatosmanoglu. 2012. A Large-scale Sentiment Analysis for Yahoo! Answers. Proc. of the 5h ACM international conf. on Web search and data mining (WSDM '12). ACM, 633-642.

[10] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann. 2011. Design lessons from the fastest Q&A site in the west. Profc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11). ACM, 2857-2866.

[11] N. Novielli, F. Calefato, and F. Lanubile. 2014. Towards Discovering the Role of Emotions in Stack Overflow. Proc. of 6th Int'l Workshop on Social Software Engineering, 33-36,

[12] F. Provost, T. Fawcett, and R. Kohavi. 1998. The case against accuracy estimation for comparing induction algorithms. Proc. of the 15th Internat'l Conf. on Machine Learning. Morgan Kaufmann, 445–453.

[13] X. Robin, N. Turck, A. Hainard, N.Tiberti, F. Lisacek, J.C. Sanchez and M. Müller. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12.

[14] M. Thelwall, K. Buckley, and G. Paltoglou. 2012. Sentiment Strength Detection for the Social Web. Journal of the American Society for Information Science and Technology, 63(1):163-173.

[15] C. Treude, O. Barzilay, M. Storey. 2011. How do programmers ask and answer questions on the web? Proc. of 33rd Internat'l Conf. on Software Engineering ICSE'11. ACM, 804-807.

[16] T.T. Ying. 2015. Mining Challenge 2015: Comparing and combining different information sources on the Stack Overflow data set. Prof. of the 12th Working Conference on Mining Software Repositories, to appear.