Towards Discovering the Role of Emotions in Stack Overflow

Nicole Novielli, Fabio Calefato, Filippo Lanubile
University of Bari
Dipartimento di Informatica
Bari, Italy
{nicole.novielli, fabio.calefato, filippo.lanubile}@uniba.it

ABSTRACT
Today, people increasingly try to solve domain-specific problems through interaction on online Question and Answer (Q&A) sites, such as Stack Overflow. The growing success of the Stack Overflow community largely depends on the will of their members to answer others' questions. Recent research has shown that the factors that push members of online communities encompass both social and technical aspects. Yet, we argue that also the emotional style of a technical question does influence the probability of promptly obtaining a satisfying answer.
In this paper, we describe the design of an empirical study aimed to investigate the role of affective lexicon on the questions posted in Stack Overflow.

Categories and Subject Descriptors
H.1.2 **[User/Machine Systems]**: Human factors

General Terms
Design, Human Factors.

Keywords
Online Q&A, Technical Forum, Sentiment Analysis, Experimental Design, Stack Overflow.

1. INTRODUCTION
The worldwide diffusion of social media has profoundly changed the way we communicate and access information. Increasingly, people try to solve domain-specific problems through interaction on online Question and Answer (Q&A) sites. The enormous success of Stack Overflow (SO), a community of over 3 million programmers asking questions (~7 millions) and providing answers (~13 millions) about software development, attests this increasing trend. Launched in 2008, Stack Overflow is now part of Stack Exchange, a fast growing network of more than 100 Q&A sites about a broad range of topics, from academic life to traveling and gaming, which originated from the success of Stack Overflow itself.

SSE'14, November 16, 2014, Hong Kong, China.

The growing success of Stack Exchange communities largely depends on the will of their members to answer others' questions. Although the factors that push members of online communities to help others are not entirely understood, they include *social aspects* (i.e., who is looking for help and their status in the community) [1] and technical aspects (i.e., what is being requested) [23]. Only recently, research has begun to investigate *linguistic factors* too, looking at how individuals write their help requests [1][16]. For instance, one of the biggest challenges in communicating through social media is to convey sentiment appropriately through text. Although display rules for emotions exist and are widely accepted for traditional face-to-face interaction, people might not be prepared for effectively dealing with the barriers of social media to non-verbal communication.

The goal of the research presented here is to understand the role of emotions in Stack Overflow. In particular, we argue that *the emotional style of a technical question influences the probability of obtaining a satisfying answer as well as the response time* (i.e., the time elapsed between the posting of a question and its accepted answer).

The popularity of Stack Overflow has made available a huge amount of interactions, written in natural language. As such, many researchers have started to analyze such data to understand the drivers of effective knowledge sharing, i.e., the main topics being discussed [4][5], which questions are answered properly [23], and which ones remain unanswered [2]. Another important issue being investigated by current research is the assessment of the quality of answers. Hart and Sarma [12] investigate how social cues and text length influence the way novice users filter and select answers on Stack Overflow. Treude et al. [23] investigated the way programmers pose and answer questions. Their preliminary findings indicate that Stack Overflow is particularly successful in replying to *how-to* questions posed by new community members. Finally, Asaduzzaman et al. [2] investigated the factors determining the success of questions and try to determine whether it is possible to predict how long a question will remain unanswered.

Unlike the increasing interests of software engineering researchers on sentiment analysis and emotion mining [7][17], existing research on online Q&A sites has not taken into consideration the potential contributions from the field of affective computing. Recent research on social media-based interaction has already demonstrated a tendency towards emotion homophily, that is, the propensity of people to share similar emotions when interacting on general-purpose social networks [21]. Expression of gratitude, urgency and reciprocity has been also demonstrated to be a factor

of success for altruistic requests on online social communities, such as Reddit.com [1]. Pletea et al. [19] performed a study on sentiment analysis of comments in GitHub discussion on security, assigning positive/negative/neutral scores to comments and pull requests from 90 GitHub projects. The findings show statistical evidence for the triggering of negative emotions among developers when dealing with security issues. In a similar study, Guzman et al. [9] provided evidence of correlation of negative sentiment with commit activity performed on Monday in 29 GitHub projects. Finally, Mitra and Gilbert [16] present a study on the cognitive and affective style of communication that happen on the Kickstarter platform in order to predict the success of crowdfunding requests. Results show the importance of the features describing the social status that the asker holds in the community as well as expressing gratitude and reciprocity in requests.

In the remainder of this paper, we first present the dataset that we intend to use in our experimental studies. Then, we describe the design of a study that we planned to answer our research question. Finally, we conclude anticipating the next steps that we are taking in our current and future work.

2. STUDY DESIGN

In this section, we describe the methodology we intend to follow to investigate how the affective lexicon of a question determines its success. Specifically, we will use a logistic regression model, which is used for predicting the outcome of a binary dependent variable based on one or more predictor (independent) variables.

In the next sections, we first detail the dataset employed, then the dependent and independent variables considered in our statistical classification model.

2.1 Stack Overflow Dataset

Stack Overflow provides data dumps of all user-generated posts, including questions and answers with tags, upvotes and downvotes, as well as information about user reputation score and badges earned. For our research, we will employ a data dump containing data from July 2008 to April 2014. Table 1 reports the amount of data available.

Table 1. Stack Overflow data dump

Item	#
Users	3,080,577
Questions	7,214,802
- with an accepted answer (successful)	4,196,125 (58%)
- without an accepted answer	2,219,421 (31%)
- with no answers (unanswered)	799,256 (11%)
Answers	12,609,623
Tags	36,923
- average tags per question	2.95
- average of tags per posts:	1.07

2.2 Dependent Variable: Defining a Successful Question

Overall, our study will analyze over 7 million of questions, the 58% of which have received one answer that has been marked by the original asker as *accepted*, i.e., the answer provides a working solution to the problem described in the question.

Therefore, as already proposed by Treude et al. [23], we define as successful a question for which an accepted answer has been provided. Hence, *Success* is the binary dependent variable in our logistic regression model, which captures the significance of the independent variables in determining the success of a question.

We are aware that such definition may be too restrictive since there are questions on Stack Overflow for which exhaustive answers are provided but none is marked as accepted. Further replication of this study will address this issue providing a broader definition successful post, considering other criteria for determining whether an exhaustive answer has been actually provided (e.g., based on the upvotes provided by other users in the community).

2.3 Independent Variables: Factors Determining the Success of a Question

Previous work has investigated the distinctive features of successful questions posed by developers in online Q&A sites. We build our set of independent variable, based on the findings of previous research on helping behavior in online communities. Other than considering affective factors (i.e., measures about the use of emotion lexicon and the overall question polarity), we include also metrics expressing other intrinsic post properties as well as metrics describing the asker's social status in the community.

2.3.1 Post Properties

To describe the intrinsic properties of a post, we consider the following set of features:

- *Title* and *Post Length*: calculated as number of words [1][2], they are also used as a metric for the preliminary automatic filtering of questions that may not meet the quality standard of Stack Overflow and, therefore, need to be moderated by the staff;
- *Code snippet*: A yes/no feature indicating whether a question embeds an excerpt of code, provided as an example by the asker. The presence of code segments provides the reader with the possibility of understanding the question with no need of further external information, being one of the possible factors of success of 'review' questions [23];
- *Topic*: The topics of questions have been used in post categorization [6][23] and in studies attempting to predict the answer probability and quality associated to a given post [2][11]. Topic has been also correlated with the probability of success of a question due to the availability of experts in the community [6]. The topic of a post on Q&A sites is usually described as a function of the tags that askers attach to their question. As for Stack Overflow, the tags are user generated and might be employed inconsistently (e.g., the authors of two posts about the same topic could use different tags). Therefore, tag clustering is fundamental to avoid the risk of both assigning different topics to related posts and introducing bias due to data sparseness (note that in our Stack Overflow data dump we observe almost 37000 distinct tags). In previous research, tag clustering has been performed using either LDA for topic modeling [5] or graph-based algorithms for community detection [6];
- *Date* and *Time*: The study of Bosu et al. [6] demonstrated that the likelihood of obtaining an answer in Stack Overflow

is also affected by the time and the day of the week. Therefore, we include as features both the question posting time (GMT hour) and the day of the week.

2.3.2 Social Factors

The users' status, expressed through reputation score and badges unlocked, resulted among the best predictors of success of a request in the study by Althoff et al. [1]. In their research, the users with higher status in the community object of the study (Reddit.com) were more likely to receive the help requested. Conversely, in the study of Treude et al. [23] questions posed by novice seems to be more frequently answered than others in Stack Overflow, probably because they are easier to answer. Either way, user reputation has been demonstrated to be correlated with the success of a post. Therefore, we intend to include it in our control variables.

Since the reputation score of users in the Stack Overflow community evolve during time, we cannot use this metric in our data dump as a measure of the status of a user. On the contrary, we need to assess the reputation of the users at the time he poses a given question in our dataset. Therefore, following the approach provided in [2], we will include in our model the following metrics:

- *Question Score and Answer Score*, computed as the difference between the upvotes and downvotes received at the time of the question, considering for all the questions and all the answer previously posted by the author of the question at hand;
- *Number of accepted answers provided by the asker*, calculated as the number of answers provided by the asker that were marked as accepted, at the time of the question;
- *Number of answers accepted by the asker*, calculated as the number of answers accepted by the asker of the question at hand, at the time of the question;
- *Number of Badges*, as provided by Stack Overflow, which awards users with silver and gold badges, according to their level of expertise; the number of badges will be used as an additional feature to assess user reputation at the time of the question.

2.3.3 Affective Factors

Affective computing is now an established discipline [18] and emotion detection is becoming a key issue in several application domain. More recently, affective computing techniques have been applied also for modeling emotion contagion in written interaction on social media [21]. Regardless of their specific application domain, the maturity reached by the techniques used in affect detection from text suggested us the possibility to include affective lexicon as an additional source of information for enriching the set of independent variables in our study.

Specifically, we will include in our model the following metrics, describing the overall polarity of the post and the use of affective lexicon in formulating questions:

- *Sentiment*: The positive/negative orientation of the text, will be used as a metric for assessing the overall polarity of a question. Sentiment will be calculated for each question in our dataset using state of the art tools for sentiment analysis. Among the envisaged tools, we consider SentiStrenght[1] and the Stanford CoreNLP Package[2], which have been employed in previous research on sentiment analysis in social computing [1][14] and perform an evaluation of the positive and negative sentiment expressed by English sentences;
- *Affective word classes:* These classes are defined in the Linguistic Inquiry and Word Count taxonomy (LIWC), developed in the scope of psycholinguistic research [22]. Features bases on word count will be included to model the use of positive and negative affective lexicons in the questions in our dataset. In the wide range of expressions of affective states, politeness (with particular significance observed for gratitude) and reciprocity (i.e., the intention to 'paying kindness forward' shown by the asker) have been demonstrated to be a strong predictor for success of altruistic requests [1]. The LIWC organizes words into psychologically meaningful categories and have been used for a wide range of psycholinguistics studies on emotions, social relationships, thinking styles, and so on [22]. Among the word classes included in the taxonomy, LIWC provides linguistic categories that draw distinctions between negative and positive emotion lexicon, which we will include in our exploratory study.

3. CONCLUSIONS

Due to its wide range of potential applications, emotion recognition is an increasingly important research area in social software engineering and, more in general, in social computing. Instead, a gap exists in literature about the role of emotions and their expression in Stack Overflow. The overall goal of our ongoing research is to investigate the role played by emotion lexicon in online Q&A websites, with a particular focus on Stack Overflow. This research is also expected to have practical implications in terms of the definition of new guidelines for practitioners and other researchers who intend to improve emotional interface design. In particular, such enhancements would enable: (i) better user experience and engagement; (ii) the development of new tools for embedding emotional intelligence into online Q&A communities to support both community members and managers.

4. ACKNOWLEDGMENTS

This work is partially funded by the E.Showcard Living Lab under the Apulian ICT Living Labs program.

5. REFERENCES

[1] Althoff, T., Danescu-Niculescu-Mizil, C., and Jurafsky, D. 2014. How to Ask for a Favor: A Case Study on the Success of Altruistic Requests. *In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM 2014)*.

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

[3] Baccianella, S., Esuli, A., and Sebastiani, F. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *In Proceedings of the Seventh Conference*

[1] http://sentistrength.wlv.ac.uk

[2] http://nlp.stanford.edu/software

on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA).

[4] Bajaj, K., Pattabiraman, K., and Mesbah, A. 2014. Mining questions asked by web developers. *In Proceedings of the 11th Working Conference on Mining Software Repositories (MSR 2014)*. ACM, New York, NY, USA, 112-121.

[5] Barua, A., Thomas, S. W., and Hassan, A. E. 2012. What are developers talking about? An analysis of topics and trends in Stack Overflow. *Empirical Software Engineering*, June 2014, Volume 19, Issue 3, pp 619-654.

[6] Bosu, A., Corley, C.S., Heaton, D., Chatterji, D., Carver, J.C., and Kraft, N.A. 2013. Building Reputation in StackOverflow: An Empirical Investigation. 2013 *In Proceedings of the 10th IEEE Working Conference on Mining Software Repositories (MSR 2013)*, 89, 92.

[7] Brooks, M., Kuksenok, K., Torkildson, M.K., Perry, D., Robinson, J.J., Scott, T.J., Anicello, O., Zukowski, A., Harris, P., and Aragon, C.R. 2013. Statistical affect detection in collaborative chat. In *Proceedings of the 2013 conference on Computer supported cooperative work* (CSCW '13). ACM, New York, NY, USA, 317-328.

[8] Casey, V. and Richardson, I. 2008. Virtual teams: understanding the impact of fear. *Softw. Process* 13, 6 (November 2008), 511-526.

[9] Guzman, E., Azócar, D., and Li, Y. 2014. Sentiment analysis of commit comments in GitHub: an empirical study. *In Proceedings of the 11th Working Conference on Mining Software Repositories (MSR 2014)*. ACM, New York, NY, USA, 352-355.

[10] Guzman, E. and Bruegge, B. 2013. Towards emotional awareness in software development teams. *In Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2013)*. ACM, New York, NY, USA, 671-674.

[11] Harper, F. M., Raban, D., Rafaeli, S., and Konstan, J. A. 2008. Predictors of answer quality in online Q&A sites. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 865-874.

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

[13] Kolakowska, A.; Landowska, A.; Szwoch, M.; Szwoch, W.; Wrobel, M.R. 2013. Emotion recognition and its application in software engineering, In Proceedings of the 6th International Conference on Human System Interaction (HIS), 532,539.

[14] Kucuktunc, O., Cambazoglu, B.B., Weber, I., and Ferhatosmanoglu, H. 2012. A large-scale sentiment analysis for Yahoo! answers. *In Proceedings of the fifth ACM international conference on Web search and data mining (WSDM '12)*. ACM, New York, NY, USA, 633-642.

[15] McCallum, D.R. and Peterson, J.L. 1982. `Computer-Based Readability Indexes. *In Proceedings of the ACM '82 Conference*, pages 44-48.

[16] Mitra, T. and Gilbert, E. 2014. The language that gets people to give: phrases that predict success on Kickstarter. *In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14)*. ACM, New York, NY, USA, 49-61.

[17] Murgia, A., Tourani, P., Adams, B., and Ortu, M. 2014. Do developers feel emotions? An exploratory analysis of emotions in software artifacts. *In Proceedings of the 11th Working Conference on Mining Software Repositories (MSR 2014)*. ACM, New York, NY, USA, 262-271.

[18] Picard, R. W. 2000. *Affective Computing*. MIT Press.

[19] Pletea, D., Vasilescu, B., and Serebrenik, A. 2014. Security and emotion: sentiment analysis of security discussions on GitHub. *In Proceedings of the 11th Working Conference on Mining Software Repositories (MSR 2014)*. ACM, New York, NY, USA, 348-351.

[20] Shah, C. and Pomerantz, J. 2010. Evaluating and predicting answer quality in community QA. *In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '10)*. ACM, New York, NY, USA, 411-418.

[21] Thelwall, M. 2010. Emotion homophily in social network site messages. First Monday, [S.l.], ISSN 13960466. Available at: <http://firstmonday.org/ojs/index.php/fm/article/view/2897/2483> Date accessed: 07 Jul. 2014. doi:10.5210/fm.v15i4.2897.

[22] Tausczik, Y. R. and Pennebaker, J. W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*

[23] Treude, C., Barzilay, O., and Storey, M. 2011. How do programmers ask and answer questions on the web? (NIER track). *In Proceedings of the 33rd International Conference on Software Engineering (ICSE '11)*. ACM, New York, NY, USA, 804-807.