Sixth International Workshop on Emotion Awareness in Software Engineering (SEmotion 2021), ICSE 2021 Workshop , May 31st, 2021, online

Emotions in Computer Vision Service Q&A

Alex Cummaudo*, Ulrike Maria Graetsch*, Maheswaree K Curumsing*,
Rajesh Vasa*, Scott Barnett*, John Grundy†
* Applied Artificial Intelligence Institute, Deakin University, Geelong, Australia
† Faculty of Information Technology, Monash University, Clayton, Australia
{ *ca, maria.graetsch, m.curumsing, rajesh.vasa, scott.barnett* }*@deakin.edu.au, john.grundy@monash.edu*

***Abstract*—Software developers are increasingly using cloud-based services that provide machine learning capabilities to implement 'intelligent' features. Studies show that incorporating machine learning into an application increases technical debt, creates data dependencies, and introduces uncertainty due to their non-deterministic behaviour. We know very little about the emotional state of software developers who have to deal with such issues; and the impacts on productivity. This paper presents a preliminary effort to better understand the emotions of developers when experiencing issues with these services with the wider goal of discovering potential service improvements. We conducted a landscape analysis of emotions found in 1,425 Stack Overflow questions about a specific and mature subset of these cloud-based services, namely those that provide computer vision techniques. To speed up the emotion identification process, we trialled an automatic approach using a pre-trained emotion classifier that was specifically trained on Stack Overflow content, EmoTxt, and manually verified its classification results. We found that the identified emotions vary for different types of questions, and a discrepancy exists between automatic and manual emotion analysis due to subjectivity.**

***Index Terms*—emotion mining, stack overflow, DevX, computer vision services, empirical study**

I. Introduction

Recent advances in artificial intelligence have provided software engineers with new opportunities to incorporate complex machine learning (ML) capabilities, such as computer vision, using cloud-based 'intelligent' web services. However, the machine-learnt behaviour of these services is non-deterministic and, given the dimensions of data used, their internal inference process is hard to reason about [1]. Recent works show that developers struggle to use these services given that they are still in a nascent stage [2], infusing machine learnt behaviour to a system often results in ongoing maintenance concerns [3]. Further, the services' documentation fails to address common issues faced [2]. Thus, developers resort to online communication—such as Stack Overflow (SO)—to ask questions about their concerns, often expressing emotions such as frustration. Negative emotions have adverse effects to productivity [4], and emotions expressed by developers online have been explored [5] including on SO [6, 7]. There is a need to better understand emotions expressed by developers when using these services; such insight could be useful in assisting cloud vendors to make improvements that would generate the most value (e.g., overall service/API design, documentation of the services, clarification in error messages).

In our recent work [2], we classified the *types of issues* developers face when using these services, specifically computer vision services, by analysing 1,425 SO questions using Beyer et al.'s taxonomy [8] (see Table I). This study extends our previous work by classifying the *types of emotions* expressed in these questions, and to understand what types of questions express the strongest emotions. This serves as an initial step to formulate a prioritised set of improvements computer vision service vendors can adopt that would bring the most value to developers (based on the types of issues developers express the strongest emotions about, thereby affecting their productivity). Motivated by existing studies exploring how emotions affect productivity [5, 9, 10], we identify the emotion(s) in each SO question (if any), and investigate whether the distribution of these emotions are similar across the various types of questions.

To achieve this goal, we opted for both *automatic* and *manual* classification approaches. Firstly, we used a pre-trained machine learnt emotion classifier, EmoTxt [6, 7, 11], trained specifically on SO posts and grounded on an emotion classification model [12]. To our knowledge, EmoTxt is the only emotion classifier trained on SO data and is well-documented for this purpose. We then triangulated the emotions detected by EmoTxt for each post against the question types we classified in [2]. Given the subjective nature of emotions, we also *manually* classified a representative sample of 300 posts using the same guidelines used to annotate the EmoTxt training dataset [6], thereby assessing overall agreement between different human raters, and manual (human) classification versus automatic (EmoTxt) classification. The three key contributions of our study are:

(i) we find the distribution of emotions differs for the type of question being asked;
(ii) our analysis of the EmoTxt results when compared with manual efforts suggests that the classification model does not generalise the computer vision service domain well;
(iii) we provide a complete replication package for future research, available at https://bit.ly/2RIGQ2N.

II. Background

Studies on the role of emotions within the workplace, including the software engineering domain, have established a correlation between emotion and productivity [4, 13]. Negative emotions impact productivity negatively, whilst positive emotions impact positively. Even though in Wrobel's study [4],

TABLE I
DESCRIPTIONS OF DIMENSIONS FROM OUR INTERPRETATION OF BEYER ET AL.'S SO QUESTION TYPE TAXONOMY.

Dimension	Our Interpretation
API usage	Issue on how to implement something using a specific component provided by the API.
Discrepancy	The questioner's *expected behaviour* of the API does not reflect the API's *actual behaviour*.
Errors	Issue regarding an error when using the API, and provides an exception and/or stack trace to help understand why it is occurring.
Review	The questioner is seeking insight from the developer community on what the best practices are using a specific API or decisions they should make given their specific situation.
Conceptual	The questioner is trying to ascertain limitations of the API and its behaviour and rectify issues in their conceptual understanding on the background of the API's functionality.
API change	Issue regarding changes in the API from a previous version.
Learning	The questioner is seeking for learning resources to self-learn further functionality in the API, and unlike discrepancy, there is no specific problem they are seeking a solution for.

anger, a negative emotion, was found to generate a motivating state to "try harder" in a subset of developers, overall, *anger* was still found to have a negative impact on productivity. In recent years, researchers have focused on identifying the emotions expressed by software engineers within communication channels such as JIRA to communicate with their peers [5, 6, 9, 10]. Most of these studies make use of one of the well established emotion classification frameworks during their emotion mining process. For example, Murgia et al. [9] and Ortu et al. [5] investigated the emotions expressed by developers within an issue tracking system, such as JIRA, by labelling issue comments and sentences written by developers using Parrott's emotion framework.

In an attempt to automate the emotion mining process, the Collab team [6, 11] extended the work done by Ortu et al. [5] by developing an emotion mining toolkit, EmoTxt [11] based on a gold standard dataset collected from 4,800 SO posts (of type questions, question comments, answers, and answer comments). 12 graduate computer science students were recruited as raters to manually annotate these 4,800 SO posts using the Shaver's emotion model (love, joy, anger, sadness, fear and surprise [12]). The work conducted by the Collab team is most relevant to our study since their focus is on identifying emotion from SO posts and their classifier is trained on a large dataset of SO posts.

III. METHODOLOGY

As discussed in Section I, the wider aims of this work are to develop a prioritised set of possible improvements to computer vision service design based on the types of issues developers express (on SO) with the strongest emotions. This study makes an initial step in that direction by first identifying *what* emotions exist (if any). We formulate four RQs:

[RQ1] What emotions, if any, exist in the language used in SO questions about computer vision services?

[RQ2] Does the classification of emotions vary for different types of SO questions?

[RQ3] What level of agreement exists between manual and *automatic classification* of emotions (using EmoTxt)?

[RQ4] What level of agreement exists between individual raters who *manually classify* emotions?

A. Dataset

This paper extends our existing work by utilising our previously curated dataset of 1,425 SO questions on four popular computer vision service providers.[1] We select computer vision services as a *concrete example* of intelligent web services due to their mature presence of these types of services available to developers (see [2] and Section VI). Each question is classified a question type per the taxonomy prescribed in Beyer et al. [8] (for reference, we provide our interpretation of this taxonomy within Table I). For further details on how this dataset was produced, we refer to the original paper [2].

After performing additional cleansing of this dataset (to remove noise), we performed *both* automatic and manual emotion classification based on Shaver et al.'s emotion taxonomy [12]. Automatic emotion detection was performed using the EmoTxt classifier, and manual classification was performed by three co-authors on a sample of 300 posts. As this was a preliminary investigation, we iteratively explored several smaller representative samples to see if any emotions could be detected and agreed upon (see Section III-D). We calculated the inter-rater reliability between EmoTxt and our manually classified questions in two ways: (i) to see the overall agreement between the three raters in applying the Shaver et al. emotions taxonomy, and (ii) to see the overall agreement with EmoTxt's classifications. Additional dataset cleansing and results from manual and automatic emotion classification are available online at https://bit.ly/2RIGQ2N.

B. Dataset Cleansing and Classifying Issue Types

As described in [2], the 1,425 questions extracted were split into 5 random samples. The first author classified the first sample of 475 questions, with three other research assistants[2] classifying the remaining 900 questions over samples of 300 posts. The remaining 50 posts were used for reliability analysis, whereby these 50 posts were classified nine times by various researchers in our group, resulting in a total of 450 classifications for the 50 posts.

Each question was classified a question issue type (as described by Table I) or, where the question was a false-positive resulting from our original search query, we flagged

[1]Google Cloud Vision, AWS Rekognition, Azure Vision, IBM Watson.
[2]Software engineers with at least two years industry experience.

the post as 'noise' and removed them from further classification. 186 posts were flagged as noise, with a total of 1,239 were successfully classified a question type. To remove duplicity resulting from the reliability analysis, we applied a 'majority rules' technique to each of these 50 posts, in which the issue type most consistent amongst the nine raters per question would win. (Given the nature of reliability analysis and assessment of subjectivity, each individual rater performed classifications in isolation, and as a result did not perform a reconciliation discussion as this would lead to changes to individual responses after-the-fact.) As an example, three raters classified a post as *API Usage*, one rater classified the same post as a *Review* question and **five** raters classified the post as *Conceptual*. Therefore, the question was classified as a *Conceptual* question. However, in four cases, there was a tie in the majority. To resolve this, we used the issue type that was most classified within the 50 posts. For example, in another question, three raters each classified the same post as *Discrepancy* and *Errors*, while the remaining three raters flagged the post as noise. In this case, the tie was resolved down to *Errors* as this classification received 72 more votes than *Discrepancy* and 88 more votes than noisy posts across all classifications made in the sample of 50 posts.

C. Automatic Emotion Classification

After all questions had been classified an issue type, we applied the method by Novielli et al. [6] on our dataset consisting of questions only. We started with a file containing the 1,239 non-noise SO questions, each with its associated question type given in Table I. We pre-processed this file by extracting the question ID and body text to meet the format requirements of the EmoTxt classifier [11]. This classifier was used as it was trained on SO posts as discussed in Section II. We ran the classifier for each emotion as this was required by EmoTxt model. This resulted in six output prediction files (one file for each emotion: *Love*, *Joy*, *Surprise*, *Sadness*, *Fear*, *Anger*), which referenced a question ID and a binary value indicating emotion presence. We then merged these emotion prediction files into an aggregate file with question text and Beyer et al.'s question type classifications that was performed in [2].

D. Manual Emotion Classification

In order to evaluate and also better understand the process used by EmoTxt to classify emotions, we randomly sampled 300 SO posts of various emotion annotations resulting from EmoTxt. Each of these 300 posts were classified by three raters (co-authors of this paper) who individually reviewed the question text against each of the six basic emotions [12] and flagged an emotion if deemed present, otherwise flagging *No Emotion* instead. Each rater reviewed each question against the guidelines provided in [6]. We then conducted reliability analysis of all three rater's results to measure the similarity in which independent raters classified each emotions against each SO post. Based on suggestions in [14], we calculated Cohen's Kappa (C_κ) [15] to measure the average inter-rater agreement *between* pairs of raters, and then Light's Kappa (L_κ) [16] to measure the *overall* agreement amongst the three raters. Results are reported in Table III. Initially, we had started with a manual classification of only 25 questions, however, this revealed strong disagreement among the three human raters. We extended the process to 150 questions and identified similar level of disagreements. We classified an additional 125 questions to conclude that the disagreements were persistent, and thus concluded manual classification at 300 posts.

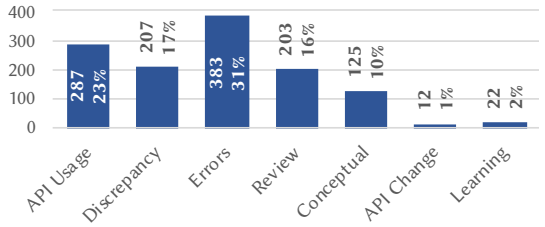

Fig. 1. Distribution of the types of questions raised.

E. Comparing Manual and Automatic Classification Methods

The next step involved comparing the ratings of the 300 SO posts that were manually annotated by the three raters against the results obtained for the same set of 300 SO posts from the EmoTxt classifier. We separated the classifications per emotion and calculated C_κ for each rater against EmoTxt, and then L_κ to measure the overall agreement. The three raters then met together to compare and discuss the ratings from the EmoTxt classifier against the manual ratings. Results are reported in Table III.

IV. Findings

Figure 1 displays the overall distribution of question types from the 1,239 posts after applying noise-filtering and majority ruling to our original 1,425 questions extracted. It is evident that developers ask issues predominantly related to API errors when using computer vision services and, additionally, how they can use the API to implement specific functionality. There are few questions related to version issues or self-learning. For further discussion into these results, we refer to [2].

TABLE II
Frequency of emotions per question type.

Question Type	Fear	Joy	Love	Sadness	Surprise	Anger	No Emotion	Total
API Usage	47	22	34	17	59	13	136	328
Discrepancy	35	12	17	7	46	20	105	242
Errors	73	34	23	21	47	23	207	428
Review	35	16	15	16	42	14	95	233
Conceptual	27	9	10	8	21	5	61	141
API Change	4	2	2	1	1	1	5	16
Learning	3	4	2	0	4	0	11	24
Total	224	99	103	70	220	76	620	1412

Table II displays the frequency of questions that were classified by EmoTxt when compared to our classification of question types. Figure 2 presents the emotion data proportionally across each type of question. In total, 792 emotions were detected within the 1,239 non-noisy posts, and 620 questions

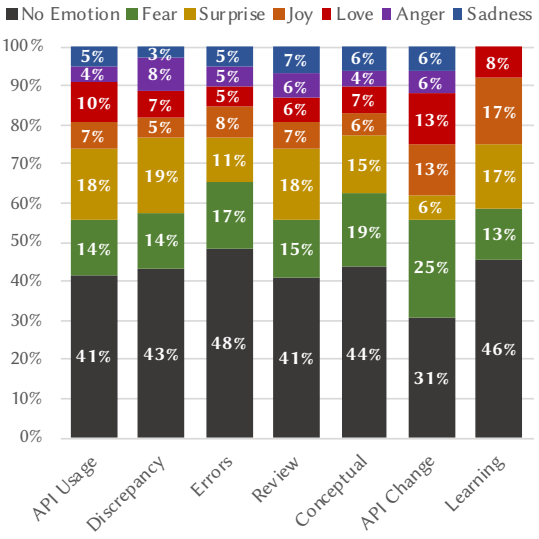

Fig. 2. Proportion of emotions per question type.

where EmoTxt predicted *No Emotion* for all the emotion classification runs. Of the 792 questions with emotion detected, 114 questions had two emotions predicted, 28 questions had three emotions detected, and one question[3] had four emotions detected (*Surprise*, *Sadness*, *Joy* and *Fear*).

***No Emotion* was the most prevalent across all question types**, which is consistent with the findings of the Collab group during the training of the EmoTxt classifier. The next highest set of emotive questions are found in the second and fourth largest samples (*Review* at 203 posts, and *API Usage* at 287 posts); therefore, higher proportions of emotion is not necessarily correlated to sample size. (Note that the broad distribution of emotions under *API Change* is not representative, as only 12 questions (1%) were classified this issue type.)

Unsurprisingly, *Discrepancy*-based questions—indicative of the frustrations developers face when the API does something unexpected—**had the highest proportion of *Anger* detected**, at 8.26%, compared to *Anger*'s mean of 4.77%. **The two highest emotions, by average, were *Fear* (μ =16.77%) and *Surprise* (μ =14.82%)**. In contrast, to our surprise, **the two least-detected emotions reported by EmoTxt were *Sadness* (μ =4.53%) and *Anger* (μ =4.77%)**. ***Joy* and *Love* were roughly the same**, and fell in between the two proportion ends, with means of 8.85% and 8.15%, respectively.

As shown in Table III, results from our reliability analysis between human raters indicated subjectivity in emotion interpretation. Guidelines of indicative strengths of agreement are provided by Landis and Koch [17], where $\kappa \leq 0.00$ is *poor* agreement, $0.00 < \kappa \leq 0.20$ is *slight* agreement and $0.20 < \kappa \leq 0.40$ is *fair* agreement. Our assessments across the 300 questions indicate slight agreement for *Love*, *Surprise*, *Sadness*, *Anger* and *No Emotion*, and fair agreement for *Joy* and *Fear*. When combining human raters and EmoTxt, the inter-rater agreement was slight across all emotions.

[3]See http://stackoverflow.com/q/55464541.

V. Discussion

RQ1: Our findings indicate that Shaver et al. [12]'s six basic emotions are detected in our questions. However, the majority of questions (42.10%) were classified by EmoTxt as *No Emotion*. It is reasonable to conclude that developers express *some* emotive language in computer vision service questions (the top three being *Surprise*, *Fear*, and *Love*), but a little under half of all questions show no emotion at all. This study set out to explore whether the emotions in these questions can highlight issues with the service. We first need to have confidence in being able to identify these emotions. EmoTxt's results show that emotional analysis of SO questions is a *partially reliable* indicator to identify developers' frustration with computer vision services. We can identify certain frustrations in *some* questions, which may lead to aspects of improvement in computer vision services. However, while some emotions may assist in identifying potential improvements for cloud vendors, not all questions will be considered due to their non-emotive language. The approach must thus be used in conjunction with other approaches to ensure that *all* questions are considered. A different classifier trained on emotive SO questions may lead to different results. We leave such exploration open to future work. Making actual service improvements is a task we leave open to future research.

RQ2: Emotions present in different types of questions classified vary, the greatest variation being in *API Change*- and *Learning*-type questions. However, these two discrepancies are a result of limited sample size; 12 and 22 questions, respectively. Thus, if we consider only the other five question types, there are distinct patterns where certain emotions are strongest. Developers express the greatest fear in *Conceptual* questions, where such emotive language likely accounts for a gap in the developer's theoretical understanding behind computer vision techniques, supporting our earlier work [2]. The greatest amount of *Surprise* and *Anger* is in *Discrepancy* questions, since the API is not behaving as the developer anticipates. Some emotions are harder to decipher. EmoTxt classified 8.15% of questions as *Love* across all of the different question types, most prevalently in the *API Usage* questions. We expected this emotion to be least expressed by developers when they encounter issues. Thus, while the type of question will entail more or fewer emotions than others, *interpreting* the reasons behind varying emotions is the more challenging factor. As it is impossible to follow-up with the authors of these questions and decipher the reasons for their emotional state, future studies with variant techniques could be guided by our results.

RQ3: Our findings in comparing manually annotated SO posts and automatic classification revealed substantial discrepancies. Table IV provides some sample questions. The subset of questions analysed by our three raters do not indicate the automatic (EmoTxt) emotion. Upon manual inspection of the text an introspection of the dataset sheds some light on this.

TABLE III
INTER-RATER AGREEMENT BETWEEN HUMANS ($R_{1..3}$) AND EMOTXT (E) AND INDICATIVE GUIDELINES OF STRENGTH.

Emotion	$\mathbf{C}_\kappa(\mathbf{R_1}, \mathbf{R_2})$	$\mathbf{C}_\kappa(\mathbf{R_1}, \mathbf{R_3})$	$\mathbf{C}_\kappa(\mathbf{R_2}, \mathbf{R_3})$	$\mathbf{L}_\kappa(\mathbf{R_{1..3}})$	$\mathbf{C}_\kappa(\mathbf{R_1}, \mathbf{E})$	$\mathbf{C}_\kappa(\mathbf{R_2}, \mathbf{E})$	$\mathbf{C}_\kappa(\mathbf{R_3}, \mathbf{E})$	$\mathbf{L}_\kappa(\mathbf{R_{1..3}}, \mathbf{E})$
Love	0.30 *Fair*	0.17 *Slight*	0.04 *Slight*	**0.17 *Slight***	0.37 *Fair*	0.27 *Fair*	0.05 *Slight*	**0.20 *Slight***
Joy	0.21 *Fair*	0.16 *Slight*	0.57 *Fair*	**0.31 *Fair***	0.1 *Slight*	0.07 *Slight*	-0.01 *Poor*	**0.18 *Slight***
Surprise	0.21 *Fair*	0.13 *Slight*	0.15 *Slight*	**0.16 *Slight***	0.17 *Slight*	0.04 *Slight*	0.06 *Slight*	**0.13 *Slight***
Sadness	0.11 *Slight*	0.05 *Slight*	0.01 *Slight*	**0.05 *Slight***	0.09 *Slight*	0.04 *Slight*	0.02 *Slight*	**0.05 *Slight***
Fear	0.19 *Slight*	0.22 *Fair*	0.36 *Fair*	**0.26 *Fair***	-0.02 *Poor*	-0.06 *Poor*	0.01 *Slight*	**0.12 *Slight***
Anger	0.19 *Slight*	0.19 *Slight*	0.07 *Slight*	**0.15 *Slight***	0.13 *Slight*	0.16 *Slight*	0.03 *Slight*	**0.13 *Slight***
No Emotion	0.30 *Fair*	0.16 *Slight*	0.09 *Slight*	**0.18 *Slight***	0.25 *Fair*	0.06 *Slight*	0.04 *Slight*	**0.15 *Slight***

TABLE IV
SAMPLE OF VARIOUS QUESTION TYPES ($[Q]$) AGAINST EMOTION(S) IDENTIFIED BY EMOTXT ($[E]$) AND THE THREE RATERS ($[R_{1..3}]$).

Question (Located at https://stackoverflow.com/q/[ID])	**Classifications**
51444352: *"I'm pretty sure I set up my IAM role appropriately (I literally attached the ComprehendFullAccess policy to the role) and the Cognito Pool was also setup appropriately (I know this because I'm also using Rekognition and it works with the IAM Role and Cognito ID Pool I created) and yet every time I try to send a request to AWS Comprehend I get the error... Any idea of what I can do in this situation?"*	$[Q]$: Errors $[E]$: Joy $[R_1]$: Surprise $[R_2]$: Surprise $[R_3]$: Anger
53117918: *"Ok so I have been stuck here for about more than a week now and I know its some dumb mistake. Just can't figure it out. I am working on a project that is available of two platforms, Android & iOS. Its sort of a facial recognition app... Is there anything I need to change? Is there any additional setup I need to do to make it work?Please let me know. Thanks."*	$[Q]$: Discrepancy $[E]$: Love, Surprise, Anger $[R_1]$: Sadness, Anger $[R_2]$: Sadness, Anger $[R_3]$: Anger
52829583: *"I was trying to make the google vision OCR regex searchable... it fails when there is the text of other languages.It's happening because I have only English characters in google vision word component as follows.As I can't include characters from all the languages, I am thinking to include the inverse of above... So where can I find ALL THE SPECIAL CHARACTERS WHICH ARE IDENTIFIED AS A SEPARATE WORD BY GOOGLE VISION? Trial and error, keep adding the special characters I find is one option. But that would be my last option."*	$[Q]$: Review $[E]$: Anger $[R_1]$: Joy, Anger $[R_2]$: Anger $[R_3]$: Surprise

For example, the first question in Table IV shows no indication of *Joy*, but EmoTxt classifies it to this emotion. Phrases like *"I'm **pretty** sure..."* could be the reason why poor classification occurred, where words like "pretty" are associated with *Joy*, albeit in a completely different context. It seems more likely the developer is experiencing a confusing situation and thus $[R_1]$ and $[R_2]$ noted *Surprise*. Similarly, in the second question presented in Table IV, EmoTxt classifies *Love*, *Surprise*, and *Anger*. It is difficult to find an element of love or appreciation elsewhere in this context beyond closing remarks: *"**Please** let me know. **Thanks**."*. Moreover, the disparity between EmoTxt and the agreed emotions between the first two reviewers shows that EmoTxt cannot detect the frustration (*Anger*) in the developer's tone, evident in their opening sentence, *"I have been **stuck here** for about more than a week and I know it is some **dumb mistake**."*. Our results indicate that further work is needed to refine EmoTxt when it is applied in new domains, like computer vision service Q&A. As highlighted by Curumsing [18], the divergence of opinions with regards to an emotion classification model proposed by theorists raises doubts to the foundations of basic emotions. Most studies of emotion mining from text use existing general purpose emotion frameworks from psychology [5, 6, 19], none tuned for the software engineering domain.

RQ4: Given the complexity and subjectivity of emotions [18], the efforts by the Collab team in automating emotions from SO posts is commendable. However, as our results have shown, agreement between a group of diverse individual raters indicates substantial subjectivity in the interpretation of emotions on SO. Without the use of reconciliation discussions to merge disparate emotional interpretations, our findings suggest that individuals will classify results based on their own personal biases. Can classification of emotions in SO be fully automated? The alternative of having human raters is expensive and time consuming. Should we therefore work towards a mid-way solution? One area of exploration is to reproduce classifications of emotions on the EmoTxt training dataset without reconciliation discussions and assess the overall reliability in the results.

VI. THREATS TO VALIDITY

Internal Validity: The *API Change* and *Learning* question types were few in sample size (only 12 and 22 questions, respectively). The emotion proportion distribution of these question types are quite different to the others. Given the low number of questions, the sample is too small to make confident assessments. Moreover, 475 of the question types were classified by a single rater; a reliability analysis on these posts would be warranted. Lastly, our classifications of Beyer et al.'s question type taxonomy was single-label; a multi-labelled approach may work better, however analysis

of results would become more complex. A multi-labelled approach would be indicative for future work.

External Validity: EmoTxt was trained on questions, answers and comments, however our dataset contained questions only. It is likely that our results may differ if we included other discussion items, however we wished to understand the emotion within developers' *questions* and classify the question based on the question classification framework by Beyer et al. [8]. Moreover, this study has only assessed frustrations within the context of a concrete domain; intelligent computer vision services. The generalisability of this study to other intelligent services, such as natural language processing services, or conventional web services, may be different. Furthermore, we only assessed four popular computer vision services; expanding the dataset to include more services, including non-English ones, would be insightful. We leave this to future work.

Construct Validity: Some posts extracted from SO were false positives. Whilst flagged for removal, we cannot guarantee that all false positives were removed. Furthermore, SO is known to have questions that are either poorly worded or poorly detailed, and developers sometimes ask questions without doing any preliminary investigation. This often results in down-voted questions. We did not remove such questions which may influence the measurement of our results.

VII. Conclusion

In prior work [2], we identified types of issues asked on Stack Overflow (SO) about four popular computer vision services.[1] Our ultimate goal is to prioritise which of these types of issues are worth addressing by cloud service vendors to effect useful improvements. To judge emotive Q&A discussion of such services we trialled a pre-trained emotion classifier trained on SO posts. This tried to determine which of these issues have the strongest emotions, since prior work has demonstrated that emotions can affect developer productivity [5, 9, 10]. We identified that EmoTxt did not classify any emotions to 42.10% of the 1,425 SO questions curated in [2]. Of the questions that did appear to express emotive language (according to EmoTxt), we found that the distributions of emotions varied according to the different types of questions posed. Given that emotions are subjective [18], three raters performed an inter-rater reliability analysis on the results of EmoTxt against a random sample of 300 posts, following the guidelines used to label data in EmoTxt [6]. Despite strictly adhering to these guidelines, we found that we could not find strong agreement between the three raters and EmoTxt nor indeed amongst the three raters themselves. Our results suggest EmoTxt classification does not generalise into new domains, such as computer vision Q&A, due to subjectivity bias.

Consistent with prior work [20], our results demonstrate that certain machine-learnt classifiers are not fully reliable for emotion classification in SO. While manual assignment of emotions is an arduous and time-consuming process, following strict guidelines (i.e., [6]) still yields subjectivity in the emotions classified. As our results highlight, applying a pre-trained emotional classifier in a new domain, such as computer vision services, will yield subjectivity issues and thus generalise poorly.

References

[1] A. Cummaudo, R. Vasa, J. Grundy, M. Abdelrazek, and A. Cain, "Losing Confidence in Quality: Unspoken Evolution of Computer Vision Services," in *Proceedings of the 35th IEEE International Conference on Software Maintenance and Evolution*. Cleveland, OH, USA: IEEE, December 2019, pp. 333–342.

[2] A. Cummaudo, R. Vasa, S. Barnett, J. Grundy, and M. Abdelrazek, "Interpreting Cloud Computer Vision Pain-Points: A Mining Study of Stack Overflow," in *Proceedings of the 42nd International Conference on Software Engineering*. Seoul, Republic of Korea: ACM, July 2020.

[3] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," in *Advances in neural information processing systems*, 2015, pp. 2503–2511.

[4] M. R. Wrobel, "Emotions in the software development process," in *2013 6th International Conference on Human System Interactions (HSI)*, 2013, pp. 518–523.

[5] M. Ortu, A. Murgia, G. Destefanis, P. Tourani, R. Tonelli, M. Marchesi, and B. Adams, "The emotional side of software developers in jira," in *Proceedings of the 13th International Conference on Mining Software Repositories*. ACM, 2016, pp. 480–483.

[6] N. Novielli, F. Calefato, and F. Lanubile, "A gold standard for emotion annotation in stack overflow," in *2018 IEEE/ACM 15th International Conference on Mining Software Repositories (MSR)*. IEEE, 2018, pp. 14–17.

[7] F. Calefato, F. Lanubile, F. Maiorano, and N. Novielli, "Sentiment polarity detection for software development," *Empirical Software Engineering*, vol. 23, no. 3, pp. 1352–1382, 2018.

[8] S. Beyer, C. Macho, M. Pinzger, and M. Di Penta, "Automatically classifying posts into question categories on stack overflow," in *the 26th Conference*. Gothenburg, Sweden: ACM, 2018, pp. 211–221.

[9] A. Murgia, P. Tourani, B. Adams, and M. Ortu, "Do developers feel emotions? An exploratory analysis of emotions in software artifacts," in *Proceedings of the 11th Working Conference on Mining Software Repositories*. Hyderabad, India: ACM, may 2014, pp. 262–271.

[10] D. Gachechiladze, F. Lanubile, N. Novielli, and A. Serebrenik, "Anger and its direction in collaborative software development," in *2017 IEEE/ACM 39th International Conference on Software Engineering: New Ideas and Emerging Technologies Results Track (ICSE-NIER)*. IEEE, 2017, pp. 11–14.

[11] F. Calefato, F. Lanubile, and N. Novielli, "EmoTxt: a toolkit for emotion recognition from text," in *Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos*. San Antonio, TX, USA: IEEE, oct 2017, pp. 79–80.

[12] P. Shaver, J. Schwartz, D. Kirson, and C. O'connor, "Emotion knowledge: further exploration of a prototype approach." *Journal of personality and social psychology*, vol. 52, no. 6, p. 1061, 1987.

[13] M. R. Wrobel, "The impact of lexicon adaptation on the emotion mining from software engineering artifacts," *IEEE Access*, vol. 8, pp. 48 742–48 751, 2020.

[14] K. A. Hallgren, "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial," *Tutorials in Quantitative Methods for Psychology*, vol. 8, no. 1, pp. 23–34, Feb. 2012.

[15] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[16] R. J. Light, "Measures of response agreement for qualitative data: Some generalizations and alternatives." *Psychological Bulletin*, vol. 76, no. 5, pp. 365–377, 1971.

[17] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–17, Mar. 1977.

[18] M. K. Curumsing, "Emotion-oriented requirements engineering," Ph.D. dissertation, PhD dissertation. Swinburne University of Technology, 2017.

[19] O. Bruna, H. Avetisyan, and J. Holub, "Emotion models for textual emotion classification," *Journal of Physics: Conference Series*, vol. 772, p. 012063, 11 2016.

[20] B. Lin, F. Zampetti, G. Bavota, M. Di Penta, M. Lanza, and R. Oliveto, "Sentiment analysis for software engineering: How far can we go?" in *Proceedings of the 40th International Conference on Software Engineering*, 2018, pp. 94–104.