Exploring the Evolution of Stack Overflow Discussions Using Sentimental Analysis on Comments

Eke, Norbert
Carleton University
Norbert.Eke@carleton.ca

Manes, Saraj Singh
Carleton University
Saraj.Manes@cmail.carleton.ca

Abstract

Stack Overflow (SO) is a popular Q&A forum for software developers, providing a large amount of discussion in form of posts and their comments. SO posts evolves with time, both in text and code snippets, so does the associated discussion with them. In this paper, we investigate the evolution of SO posts with respect to SO discussions, a factor usually ignored in techniques aimed to find relevance of a post for particular objective. To accomplish our goal, we mine SOTorrent data set that provides version history of posts and comments with time line. We then study the characteristics of discussions in form of comments with respect to evolution time line of post. Our results demonstrate that on average sentimental trend favors positive sentiment as posts becomes more stable with time, characterizing more approval from SO community in comment section.

1 Introduction

Stack Overflow (SO) is the most popular question answering website for software developers, providing a large amount of code snippets and free-form text on a wide variety of topics. In the latest public data dump from December 9th 2018 SO listed over 42 million posts from almost 10 million registered users. Similar to other software artifacts such as source code files and documentation, text and code snippets on SO evolve over time. An example of this is when the SO community fixes bugs in code snippets, clarifies questions and answers, and updates documentation to match new API versions. To analyze how SO posts evolve (Baltes et al., 2018) built SOTorrent, an open data set providing access to version histories of SO content at two separate levels: whole posts and individual text or code blocks (**?**).

Since the existence of SO starting in 2008, a total of 13.9 million SO posts have been edited after their creation. 19,708 of these posts have been edited even more than ten times. 300 million software developers and engineers visit Stack Overflow monthly (**?**). This number shows the scale of interaction happening on this platform. When a question is asked or answered, most of the discussion and interaction related to that topic happens in the form of comments. Comments can be associated with a question or an answer. These comments provide rich source of natural language text (mainly in English) to study developers' attitude towards a topic. In the submission paper (**?**) the authors performed some initial analysis on this database. Baltes et al. claimed that out of all posts on SO, 38.6% have been edited after their creation. Furthermore, the authors of the paper argued that all edited posts are very rich in comments, having a large number of comments compared to non-edited posts. They inferred that these comments lead to the edit of the post. As part of this project we would like to perform very specific sentimental analysis on these text rich comments of edited post to argue about nature of these comments.

In (**?**) the authors explored a question's preferred answer ranking as a binary classification task. Their classifiers performed well in general but failed when sentiment of comment were negative because purposed classifier did not take comment sentiment into consideration. Thus, we believe that to build better algorithms for question-answer ranking or best answer selection on SO, a detailed analysis of discussions related to an answer, is important. Rather than just considering content of answer or post, surrounding discussions should also be analyzed and considered for building such techniques.

As part of this research project, we are focusing

on the effect of SO discussions in form of comments and their effect on post/answer evolution with time. In analysis of discussions we are taking a comment as an atomic unit for sentiment analysis. We are creating an edit time line for each post, as we want to focus on the evolution of posts in terms of edits, since the discussions are associated with them.

1.1 Summary of Contributions

The contributions of this project are the following:

- Establishing and confirming evolution of posts on SO.
- Establishing that sentiment of SO discussions play a role in evolution of posts.
- Discovering that change of sentiment can be captured by sentiment valence plots.
- Analyzing tag specific sentiment trends of SO discussions and discovering trend patterns in discussions with most frequent tags.

The rest of the paper moving forward will contain related works in section 2, then the methodology in section 3, including data collection (3.1), text processing (3.2), sentiment valence analysis (3.3), sentiment polarity trend analysis (3.4) and tag specific sentiment polarity trend analysis (3.5). Results and answers to research questions can be found in section 4, then threat to validity will follow in section 5, and lastly future work and conclusion will conclude the paper in sections 6 and 7.

2 Related work

A signification amount of research work has been done on sentiment analysis of Stack Overflow. (**?**), (**?**) are widely accepted and cited works in regard to sentiment analysis of Stack Overflow discussions. While they analyze SO comments independently, they fail to capture their relation and effect with SO posts. In our published work (**?**), we analyzed the evolution of posts in terms of code snippets. This project is further extending on the evolution theme, however here we are looking at post evolution from a different angle.

2.1 Objectives

As part of this project we would like to answer the following four research questions:

Number of Posts	684
Total Number of comments	14283
Average number of comments per post	21
Total number of words	499161
Vocabulary Size	49638

Table 1: Characteristics of the final data set

RQ1: How often do SO posts get edited ? Do SO posts evolve ?

RQ2: For all edited posts on SO, what is the overall sentiment of comments with respect to edit time line? Does the overall sentiment in discussions improve with edits?

RQ3: Assuming quality of SO answers improves with edits, after how many edits does the SO answer improve/stabilize? (Assumption made: SO post answers improve when negative sentiment comments disappear)

RQ4: How does sentiment of SO discussions vary with tags? Is sentiment trend similar across different tags?

3 Methodology

3.1 Data Collection

Data selected for this project is a subset of our data set from (**?**). In our previous work we selected SO posts which were mostly refereed in open source GitHub projects. This consisted of around 21 thousand SO Posts. Among these posts were both question and answer posts. To further trim down the size of data set we imposed 2 conditions on this data set:

1. Post should be an accepted answer
2. Post should have been edited at least 5 times.
3. Post should have at least 10 comments.

Resulting data set had 684 posts. The unit of analysis for this project is a comment and edit time line of a post, rather than content of a post. Table 1 contains the characteristics of the final data set for processing.

3.2 Text Processing

The textual data within the data set comes in the form of comments and edits to post answers. These comments and edits may contain user name mentions, source code, variable names, hyperlinks, function names and API calls or references

to documentation. For this specific reason only very limited amount of text pre-processing was done. Lemmatization, stop word removal, stripping of tags, numeric values and punctuation was avoided in order to keep the software and source code related format and information intact.

However, removal of multiple white spaces, user name mentions and HTML links was performed to carefully clean up the text within the comments and edits of SO posts. Other data transformations included converting the date and time of creation and editing of a SO post from a string date to a *datetime* object in Python, then sort the data set by ascending *Post ID*, *Creation Date* and *Edit Date*.

3.3 Sentiment Valence Analysis

Performing sentiment analysis on comments and edits of an SO posts' discussion had to be different than just looking at positive and negative sentences. The goal was to identify the times when there was a change in sentiment within the discussion on a post, therefore performing traditional sentiment analysis by classifying each sentence's sentiment polarity was not the perfect solution. The exploration of different kinds of sentiment analyzers were considered with the main goal in mind to be able to see how sentiment changes over time, creating a sentiment time line for the discussion within a SO post. The R package named Syuzhet performs exactly such kind of analysis by looking at the "latent structure of narrative by means of sentiment analysis" (**?**). The author of the package described that "it reveals the sentiment shifts within a sequence of sentences [or SO post discussion thread] by extracting sentiments and creating sentiment-based plot arcs from text" (**?**). The Syuzhet package[1] contains four sentiment dictionaries and provides a method for sentiment extraction developed by the NLP group at Stanford. Most methods within this library depends on the coreNLP package[2]. For our research purposes the default "Syuzhet" lexicon was used, which was developed in the Nebraska Literary Lab.

Using this R package the creation of a sentiment time line for the discussion in each analyzed SO post was made possible. The Syuzhet package extracts sentiments and calculates a sentiment valence score at each point on the time line generating a sentiment valence plot. Using these sentiment valence plots all the sentiment changes over time are visually observable, while the individual sentiment valence scores are analyzable.

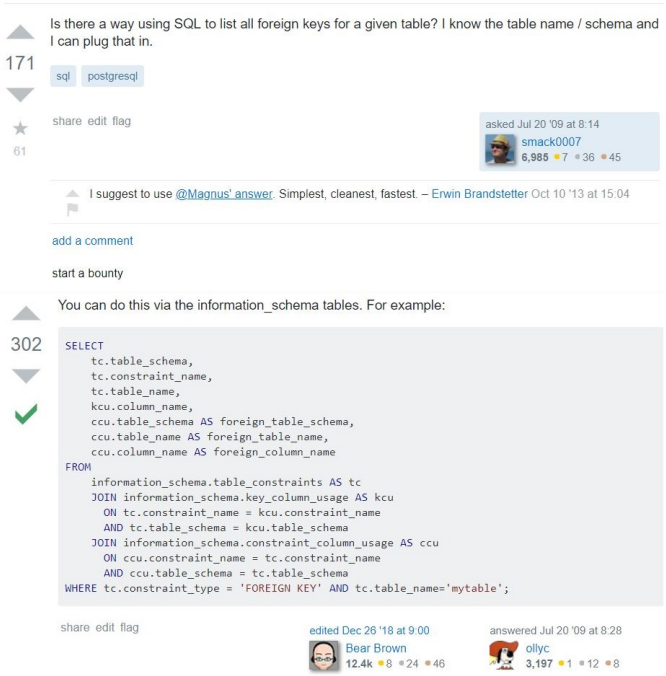


Figure 1: Sample question and its accepted answer on Stack Overflow

Figure 1 shows an example of a question and its accepted answer on Stack Overflow. While most of the SO community only looks at the accepted answer, it is not well known that the accepted answers go through an evolution by constantly being edited by the answerer and the SO community. In figure 1 it can be seen that the question was asked on July 20th 2009, and it has been going through edits for 9 years, with last edit happening in 2018. This phenomenon is referred to as the evolution of SO posts by the MSR (Mining Software Repositories) community.

Figure 2b shows part of the discussion about the question and accepted answer shown in figure 1. Figure 2a is the sentiment valence plot generated by the Syuzhet package, which serves as the visual representation of changes in sentiment within the discussion. The SO community agreeing or disagreeing with an answer can be speculated by looking at the sentiment trends on the sentiment valence plot. In figure 2a the x-axis is a time line of all comments and edits happening in discussion

[1] https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html

[2] https://stanfordnlp.github.io/CoreNLP/

.png

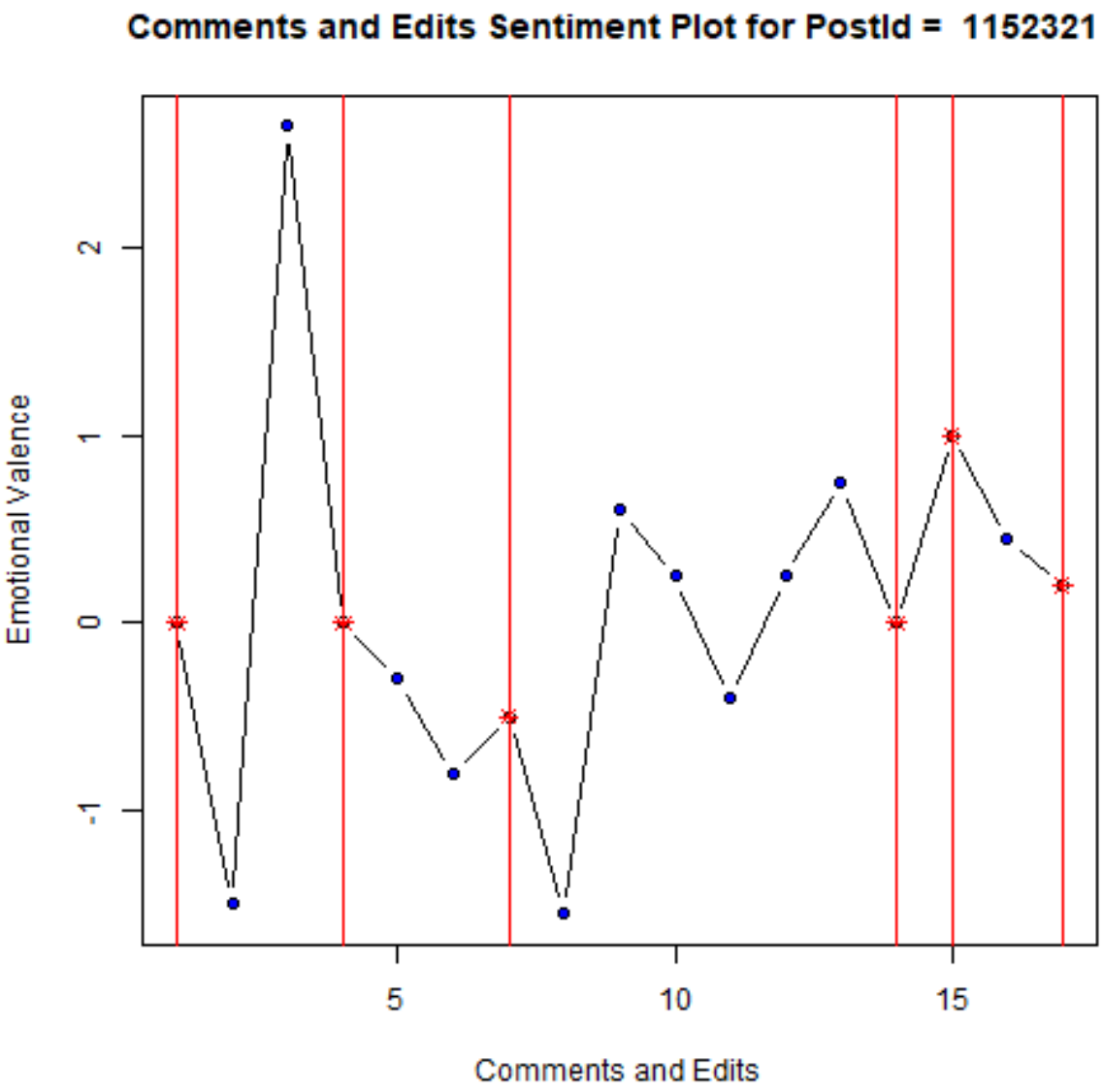
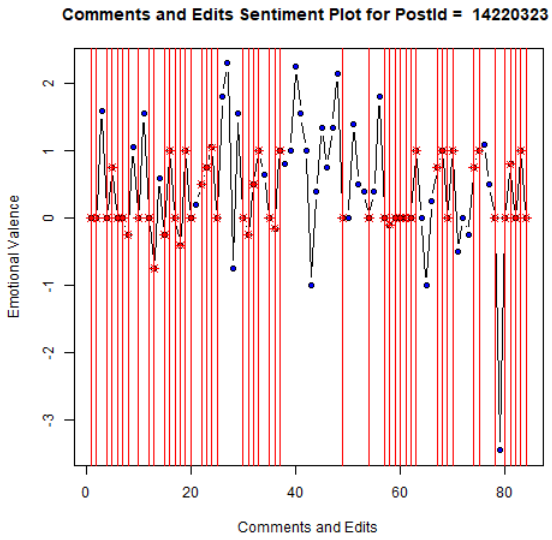


Figure 2: Discussion about the accepted answer on Stack Overflow and visual representation of the sentiment trend within the same discussion using a sentiment valence plot



.png

Figure 3: Sample sentiment valence plot generated by the Syuzhet package

about an SO post. The y-axis is a sentiment valence score. The higher values indicate positive sentiment in the discussion, while lower values indicate negative sentiment in the discussions. Blue dots represent individual comments, while red vertical lines indicate that there has been an edit on the post at that specific time.

Figure 3 shows one more example of a sentiment valence plot. It is not straightforward to interpret and make conclusions about the sentiment trend in most SO post discussions, as there are too many changes in sentiment trend, while the large number of edits and comments between the edits clutter the plot. An overall sentiment polarity trend analysis (section 3.4) on all analyzed posts needs to be performed in order to get the insights and be able to quantify overall sentiment trend from all SO post discussions.

3.4 Sentiment Polarity Trend Analysis

In order to see overall post-wide sentiment polarity trends, the cumulative sentiment valence score of each post was calculated. Table 2 shows aggregated relative and absolute frequency counts of the overall polarity of SO posts after the first 10 edits. This table can be interpreted as out of 684 analyzed posts, after the first edit 598 posts contained an overall positive sentiment valence score, while 85 posts had overall negative valence scores, cumulatively. Based on table 2 one can see that after each edit the cumulative sentiment valence scores

Edit Count	Pos. Freq.	Neg. Freq.	Pos. Percent	Neg. Percent
1	598	85	87.55	12.45
2	593	90	86.82	13.18
3	594	76	88.66	11.34
4	573	63	90.09	9.91
5	538	51	91.34	8.66
6	491	34	93.52	6.48
7	435	23	94.98	5.02
8	372	19	95.14	4.86
9	323	13	96.13	3.87
10	272	10	96.45	3.55

Table 2: Aggregated relative and absolute frequency counts of the overall polarity of SO posts after 10 edits

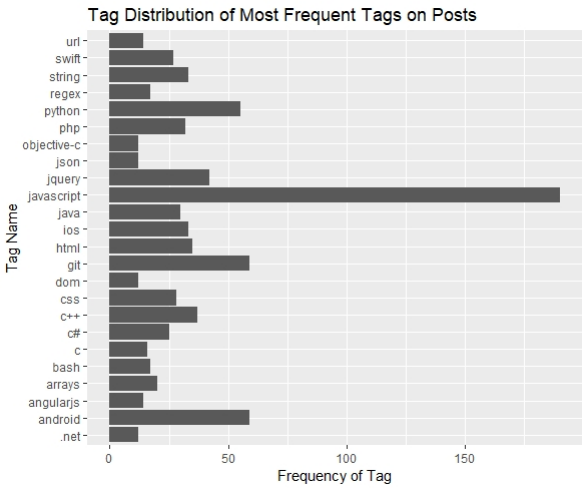


Figure 4: Distribution of tag frequencies within the analyzed posts

are mostly positive, and as the number of edits increase, the percentage of positive discussions in SO posts increase.

3.5 Tag Specific Sentiment Polarity Trend Analysis

One of the research questions explores how does sentiment of SO discussions vary with tags? Is sentiment trend similar across different tags? To look at tag specific sentiment polarity trends, tag name data had to be linked to the posts. After doing so, one can see in figure 4 the distribution of tag frequencies within the 68 analyzed posts. The posts belonging to the 5 most frequent tags **(Javascript, Python, Android, Git and jQuery)** and a couple less frequent tags **(HTML, C and C#)** went through their own tag specific sentiment

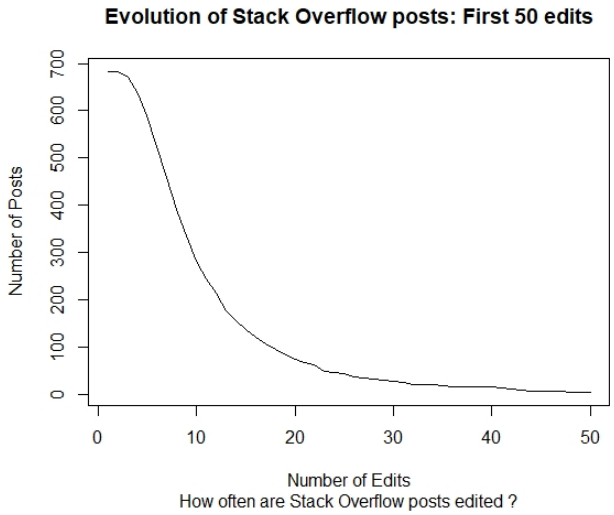


Figure 5: Evolution of SO posts: Graph of Number of Posts Edited vs. Number of Edits for the First 50 Edits

polarity trend analysis to get an insight whether or not the overall sentiment trend is similar across different tags.

4 Results

4.1 RQ1: How often do SO posts get edited ? Do SO posts evolve ?

Posts on Stack Overflow, like other software artifacts evolve with time. Their evolution can be defined in number of times they are edited. Figure 5 shows number edits and number of posts for our chosen data set. Number of posts almost exponentially decreases with number of edits. In our sample size of 684 minimum number of edits is 5, since this is our selection criteria, but maximum number of edits for a post goes as high as 768. Although our sample size is small, the same trends follow on the entire SOTorrent data set. Two consecutive edits along with time of their occurrence defines an interval for further analysis to capture trend of discussion in comment section of answer post. This interval is further utilized in answering the remaining research questions. Figure 5 also emphasizes that number of such intervals is very significant even in our small subset thus providing very rich data set to work on.

4.2 RQ2: For all edited posts on SO, what is the overall sentiment of comments with respect to edit time line? Does the overall sentiment in discussions improve with edits?

The partial results from the overall sentiment polarity trend analysis are shown in table 2. This table shows that overall the cumulative sentiment valence scores are positive, but when graphed, even more conclusions can be drawn. Figure 6a is a stacked bar chart containing the frequency of overall sentiment polarity of posts, while figure 6b is a visualization of the same data, but with a percentage ratio of the overall sentiment polarity of posts, both figures showing data after each of the first 20 edits. It is worth noting the interesting distribution of overall negative sentiment polarity trends within the first 20 edits. About 12% of posts have overall negative sentiment polarity score after the first edit, then this number increases to 13%, then slowing decreases gradually until after the 11th edit, at which it is 3%. After the 11th edit the percentage of overall negative sentiment polarity trends bounces up and down between 5% and 4%, then only after the 19th edit it goes down to about 1%, then finally the overall sentiment polarity trends converge to 100% positive polarity after the 24th edit.

Our findings shows that eventually as time progresses, sentiment trends stabilize and converge towards positive polarity. At the time of positive convergence in sentiment polarity most of the community agrees on the answer, which went through multiple iterations of editing, suggesting that the final answer has evolved enough. These results also suggest that the sentiment of comments on an answer leads to more editing of the answer. The more negative the overall sentiment polarity trend is, the more likely there will be an edit to the answer.

4.3 RQ3: Assuming quality of SO answers improves with edits, after how many edits does the SO answer improve/stabilize ?

Assuming negative comments implicate dissatisfaction or discontent in a comment about the corresponding answer post, figure 6 shows that such discontent in the comment discussion section dies down as number of edits increase. Although on large, such decrease in negative sentiment is strict in nature, a small reverse trend is observed in the

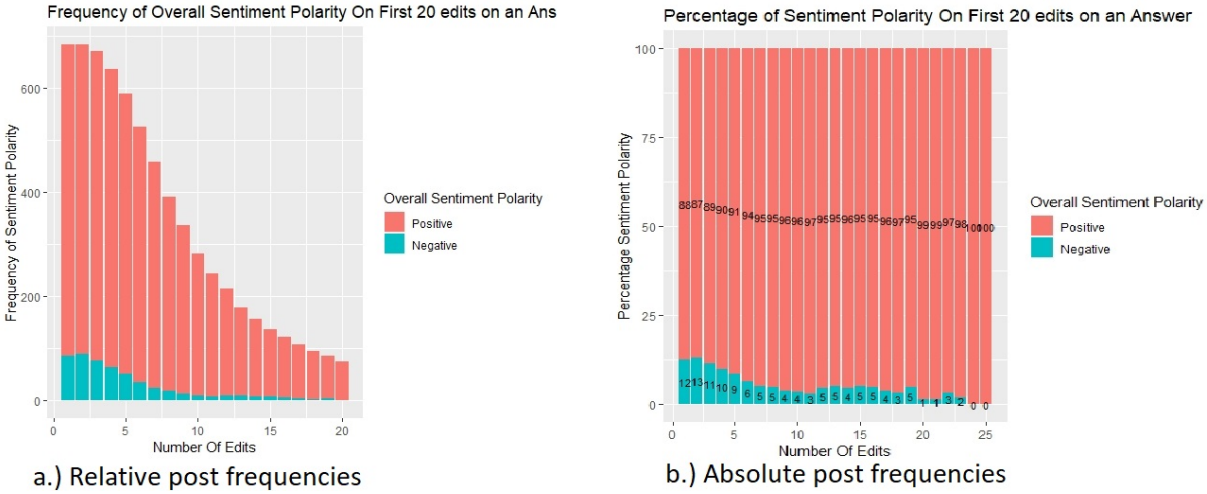


Figure 6: Frequency and percentage of overall sentiment polarity trend results for first 20 edits of all posts

first few edits, indicating enough rejection and disapproval towards the answer, close to its creation time. This trend gradually decreases, while positive sentiment increases. Further, on average we see that such negative sentiment score goes to 0 after the 20th edit, among possible 700 edits.

4.4 RQ4: How does sentiment of SO discussions vary with tags? Is sentiment trend similar across different tags?

In order to find out whether or not sentiment trends is similar across posts with different tags tag specific sentiment polarity trend analysis was performed. Tables 3 and 4 show results for the above mentioned analysis performed on posts with top 5 most frequent tags, respectively posts with 3 less frequent tags. Each cell within these two tables shows the percentage ratio between overall positive and negative sentiment polarity trends for a specific tag (found in the column header), after a specific edit number, and only the first 10 edits being displayed for convenience. In table 3 it can be see that all posts with frequent tags on average have around the same sentiment polarity ratios of positive to negative. The difference between post with different most frequent tags is the number of edits they need for the positive polarity ratio reach a high enough value, then eventually converge to 100%. If one contrasts the sentiment polarity ratios of the most frequent tags (table 3) with the some of the less frequent tags (table 3), the conclusion is that some of less frequent tags converge to 100% positive sentiment polarity faster, sometimes even after less than 10 edits, while more popular tags tend to need more rounds of editing. This observation makes sense, as the more popular a tag is, it is expected to have more discussion and more following, thus the SO community tends to debate those questions with popular tags for longer time.

To show that even the posts with popular (more frequent) tags converge to 100% positive sentiment polarity, figure 7 visualizes tag specific sentiment polarity trend results for the first 50 edits. In this figure it can be seen that as number of edits increase, all posts with popular tags converge to 100% positive sentiment polarity. For posts with some popular tag it takes longer, as there is more debate in the discussion, like posts with *Git* tag. It is interesting to note that posts with Javascript, Git and Android tags have the similar patterns of positive sentiment polarity (gradually increasing until reaching "convergence"), while posts with Python tag have quite a lot of negative sentiment polarity after the first 15 edits, then rapidly "converging" to 100% positive sentiment polarity. Within the first category of patterns, posts with Git and Android tags are even more similar, as they both reach 100% positive sentiment polarity "convergence", then destabilize for a few edits with more negative sentiment polarity, then jump back up to 100% positive sentiment polarity "convergence".

To answer the research question, generally it can be said that sentiment trend similar across different tags, but further research is needed to find well defined subgroups of multiple tags having very similar overall sentiment polarity trends.

Edit Count	Javascript sentiment ratio (pos. - neg. %)	Python sentiment ratio (pos. - neg. %)	Git sentiment ratio (pos. - neg. %)	Android sentiment ratio (pos. - neg. %)	jQuery sentiment ratio (pos. - neg. %)
1	88.9 - 11.1	92.9 - 7.1	83.3 - 16.7	88.5 - 11.5	90.2 - 9.8
2	87.3 - 12.7	94.6 - 5.4	78.3 - 21.7	91.8 - 8.2	90.2 - 9.8
3	91.5 - 8.5	90.7 - 9.3	80 - 20	94.9 - 5.1	92.5 - 7.5
4	94.9 - 5.1	89.6 - 10.4	87.7 - 12.3	94.6 - 5.4	97.3 - 2.7
5	95.7 - 4.3	88.9 - 11.1	88.9 - 11.1	97.9 - 2.1	93.8 - 6.3
6	94.6 - 5.4	90.5 - 9.5	89.6 - 10.4	97.5 - 2.5	100 - 0
7	97.1 - 2.9	89.7 - 10.3	92.9 - 7.1	97.1 - 2.9	100 - 0
8	97.4 - 2.6	87.1 - 12.9	94.4 - 5.6	96 - 4	100 - 0
9	98 - 2	92 - 8	96.9 - 3.1	95.5 - 4.5	100 - 0
10	96.5 - 3.5	90.5 - 9.5	100 - 0	100 - 0	100 - 0

Table 3: Tag specific overall sentiment polarity trend results after each one of the first 10 edits for posts with top 4 most frequent tags

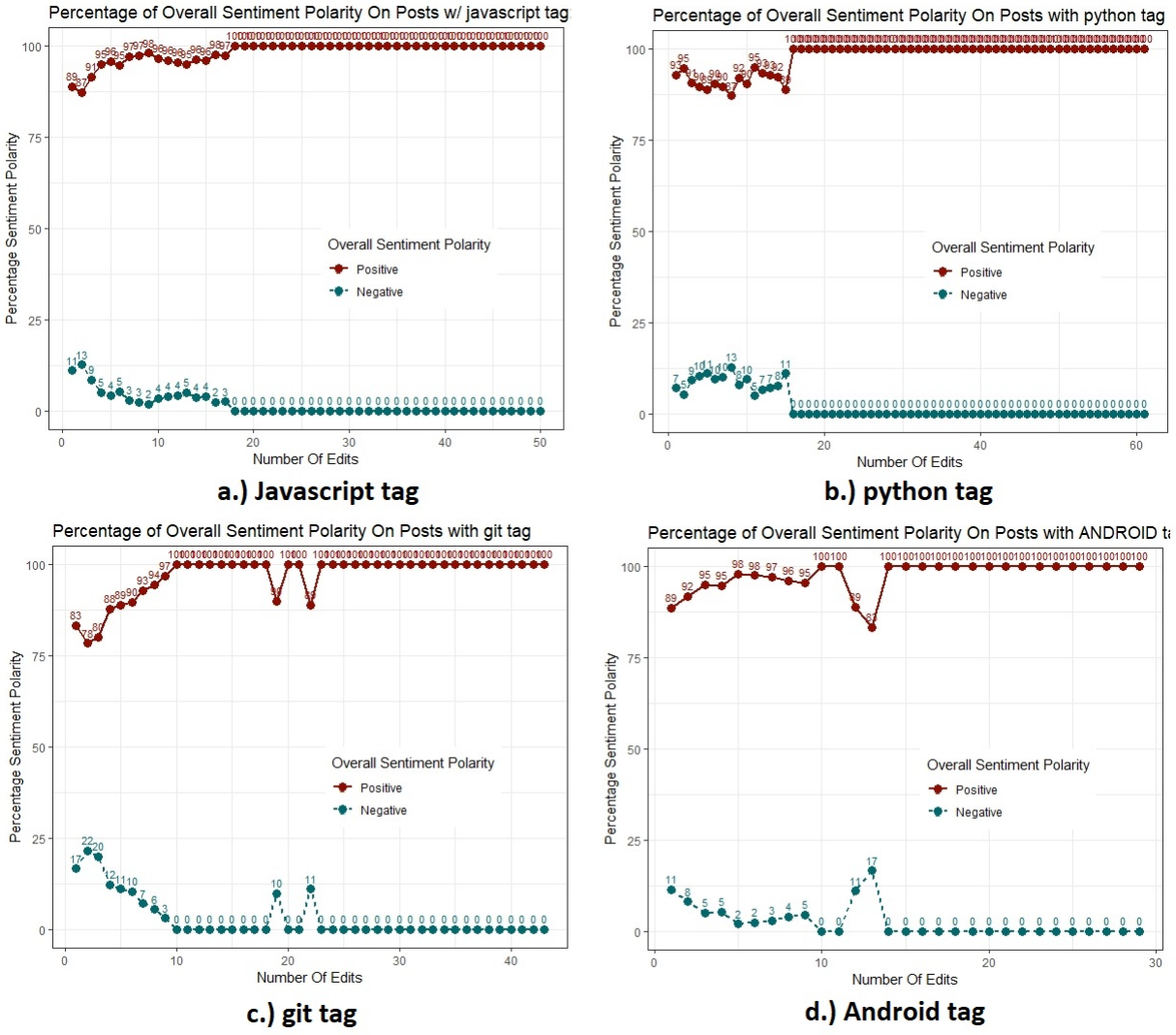


Figure 7: Visualization of tag specific sentiment polarity trend results after each one of the first 50 edits

Edit Num	HTML sent. ratio (pos-neg)	C sent. ratio (pos-neg)	C# sent. ratio (pos-neg)
1	86.4 - 13.6	75 - 25	76 - 24
2	90.9 - 9.1	81.3 - 18.8	80 - 20
3	92.9 - 7.1	81.3 - 18.8	84 - 16
4	92.5 - 7.5	80 - 20	91.3 - 8.7
5	91.9 - 8.1	85.7 - 14.3	90 - 10
6	93.5 - 6.5	100 - 0	92.9 - 7.1
7	96.4 - 3.6	100 - 0	87.5 - 12.5
8	95.7 - 4.3	100 - 0	83.3 - 16.7
9	94.4 - 5.6	100 - 0	100 - 0
10	92.9 - 7.1	100 - 0	100 - 0

Table 4: Tag specific overall sentiment polarity trend results after each one of the first 10 edits for posts with 3 less frequent tags

5 Threat to Validity

Although sample size of our data set is small, applicability of our results for Stack Overflow posts at large might seems biased. We believe that since the original data set in (**?**) was chosen at random and only considering most refereed SO Posts from SOTorrent, results should uniformly scale and should be applicable to most popular SO posts in general.

Throughout this project we have treated the training process behind the Syuzhet package [3] as a black-box function. We are not aware how the tool is assigning sentiment valence scores, but the authors of the tool have validated it on both short and longer text, thus the score calculations and learning process must still be valid.

Another threat to validity could be that the tag specific analysis' sample size might not be large enough for a stronger, more meaningful conclusion. We believe this is not a concern, as we are only looking at patterns and trends within a few popular tags.

6 Future Works

In this analysis we have ignored the time when a post is accepted as accepted time stamp was not available for all posts. But an high level analysis states that edits happens even after an post is accepted and even when overall sentiment score of related discussion is negative. This indicated that a post is prematurely accepted by the asker. We would like to demonstrate this on larger data set thus process to accept the post should be more democratized rather than by one single user and should be done once post is stabilized and discussions have more positive response from community.

Knowing if a post is reliable/stable (unlikely to be edited in future) is important for software engineers, as they need to know whether or not they can trust the answer. In the future it can be explored how to detected if a post will likely be edited in the future? Other future works could include building a machine learning model to predict the edit number where the post's answer is reliable, thus has stabilized. Lastly, another idea could be to come up with a metric and quantify post quality, then use this value to judge post reliability.

7 Conclusion

As a post evolves with time, so does the related comments section providing rich context for approval and disapproval towards answer posts. As posts get edited more with time, the SO community provides more supportive and positive responses indicating more agreement on the answer. This work will motivate further research in post reliability prediction and evolution of code snippets, edits and posts on Stack Overflow.

8 Acknowledgements

We would like to acknowledge the great advice received from our supervisor, Prof. Olga Baysal. We would also like to thank Prof. Diana Inkpen for advising us on the natural language processing aspects. Lastly, we appreciate Sebastian Baltes' help with technical details with SOTorrent.

References

Sebastian Baltes, Lorik Dumani, Christoph Treude, and Stephan Diehl. 2018. Sotorrent: Reconstructing and analyzing the evolution of stack overflow posts. In *Proceedings of the 15th International Conference on Mining Software Repositories*, pages 319–330. ACM.

[3] https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html