Unraveling Sentiment Threads: An Analysis of Comment Sentiment and User Participation in Scratch Project Creation

Investigating the Impact of Comment Sentiment on the Creator's Activity on a Social Coding Platform

Gert-Jan Schaap

Supervisor(s): Fenia Aivaloglou, Sole Pera

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 26, 2024

Name of the student: Gert-Jan Schaap
Final project course: CSE3000 Research Project
Thesis committee: Fenia Aivaloglou, Sole Pera, Jorge Martinez Castaneda

Abstract

Although existing work demonstrates that the usage of Scratch strengthens programming skills, little is known about the effect of children encouraging or criticizing each other when programming on the development of their programming skills. To address this gap, we conducted a data analysis to reveal the correlation (or lack thereof) between the sentiment of comments on a Scratch project and the creator's inclination to continue producing projects.

In particular, using a known dataset of Scratch projects, which we enriched with new projects, we examined several independent variables that capture sentiment on comments from different perspectives, e.g. the absolute or relative number of comments with a particular sentiment score that a user receives, or the ratio of comments that fall in a certain category. We also examined several dependent variables that capture the creator's inclination to continue producing projects, e.g. the number of projects they create or the total time that they are active on the platform.

The results of our experiments reveal that the absolute and relative number of comments with a specific sentiment score given to a user have a low correlation to the time that a user is active on the platform. Additionally, the ratio of the number of positive and negative comments over the total number of comments given to a user has a low correlation to the time that a user is active on the platform. Finally, the number of comments that a user receives has a low correlation to the number of projects a user creates.

1 Introduction

Scratch is an online platform that was created in 2007 and has been growing since then to become one of the largest coding communities for children in the world. The mode value for the age of new Scratchers (people who joined the Scratch platform) is 12, and there is a peak in the age distribution of new Scratchers around this value, which means that most of the users are in fact children [12]. In Scratch, children have the opportunity to learn programming concepts and to practice computational thinking visually via a block-based, colorful programming language. They can do this while meanwhile experimenting with their own creativity to create different kinds of application like games and animations.

Besides this aspect of learning to program, Graßl and Fraser have shown that Scratch also fulfills the role of a social network. This can be seen, for example, by the fact that socially relevant events such as COVID-19 or the Black Lives Matter movement are reflected in Scratch to some extent [3].

On social networks in general, the messages that are posted on that network usually impact the children that are using it in some extent [5]. In the case of Scratch specifically, this impact might be positive, since Fields et al. describe that many comments contain encouragements, suggestions and tips [2]. On the other hand, there are some Scratch users that expose societal themes like racism and bullying in their Scratch projects. Richard and Kafai noticed that those users received negative and disparaging comments on their projects, and that those users seem to have abandoned their profiles after some time or seem to be no longer active on Scratch under their known profile [9].

In addition to stories from previous research, it is also possible to find projects on the Scratch website where children explain the impact that comments made on them. For example, a user called "LooneyTooney" created a project called "BULLY a short REAL story." In that project, he mentions somebody commenting "wow. just wasted a minute of my life whatching this. you really suck at this." and the impact that that comment made on him [1].

These observations concerning these people leaving the Scratch platform raise interesting questions like: Did these users leave Scratch because of the disparaging comments they received? And, if those comments have really arised from ethnic or cultural racism: Could young girls also receive more disparaging comments than young boys due to their gender? More research would be needed to answer these questions. Of course, it is clear that it would be unjust for girls to receive more disparaging comments than boys, because when taught equally, they have a comparable coding proficiency [19]. In 2021, it was reported that only about 25% of the computer scientists and 15% of the engineers in the United States are women [7]. Also, research has shown that the usage of Scratch in the teaching of programming to students, has significantly increased students' overall motivation and examination performance [18]. However, if it were true that girls would receive more disparaging comments than boys, and if it were true that disparaging comments cause a user to stop using Scratch, then girls would also probably lose some motivation for computer science and maybe even stop learning to program at all [18]. This would have the effect that girls would develop less in their use of programming concepts and in the development of their computational thinking skills [20], which could even further strengthen the underrepresentation of women in computer science careers.

The reasoning above stands or falls with the question of whether receiving negative and disparaging comments on their projects influence a child to leave the Scratch platform. This could signify the importance of encouraging children to learn to program. This thinking process led to our research question:

To what extent is there a relation between the sentiment of comments on a Scratch project and the creator's inclination to continue producing projects?

We answer this research question by performing data analysis on seven different combinations of variables relating to the sentiment of comments and the creator's inclination to continue producing projects. On each combination of variables, we calculate corresponding correlation coefficients and we draw conclusions based on those.

Our results show that, depending on the chosen variables, the correlation is either low or not present at all. This makes

[1] https://scratch.mit.edu/projects/68436216/

the effect of sentiment of comments on a Scratch project on the creator's inclination to continue producing projects limited. However, it must be noted that some limitations have occurred during the data analysis, which are discussed in section 6.

Section 2 contains some work from previous research that we have used to make decisions. Section 3 explains the different iterations of data analysis in detail. It also explains how the data for this research has been retrieved. Section 4 demonstrates the results from the iterations of data analysis. Section 5 contains some ethical considerations that we took into account for this research. Finally, section 6 gives a discussion on the reliability of this research and section 7 provides the conclusions of this research.

2 Related Work

Velasquez et al. have shown that project comments on Scratch are often richer in language than other comments [17]. This is good for our research because richer language makes it easier for a sentiment analysis tool to recognize words, probably improving the quality of predictions.

Sentiment analysis in other datasets of Scratch comments has revealed that they are mostly positive [3; 2; 16]. Graßl and Fraser found, using VADER [4] as a sentiment analysis tool, that the tone in the comments on the Scratch platform is mostly positive. Fields et al. found that 72% of the comments on Scratch have a positive tone, while 14% of the comments on Scratch have a negative tone. Van der Ende, on the other hand, found different results; they found that only about 43% of the comments have a positive tone. Because all of this research demonstrates that there is a large number of positive comments compared to negative comments, this shows the importance for us to have a large dataset that contains enough negative comments, next to the positive comments.

Van der Ende has also published a dataset that is useful for our research. Their dataset contains comments and their sentiment analysis scores for 200.000 projects that are created in the first months of 2019 [16].

Richard and Kafai have done research on diversity in the Scratch community. Their research demonstrates the relevance of our research, since it made the observation that there might be a connection between negative comments and races [9]. Wen et al. also underline the importance, since they reveal how the use of Scratch in programming learning significantly increased students' motivation and performance [18]. The research of Richard and Kafai and the research of Wen et al. could place our research in a broader perspective of what will happen when people stop using Scratch because of negative comments.

3 Methodology

This section describes the methodology to answer the research question. Section 3.1 describes the choice we made between sentiment analysis and emotion detection. Section 3.2 describes the datasets (and their generation) that we used and created for this research. Section 3.3 describes how the relevant variables were selected. Finally, section 3.4 describes the data explorations that we have done in detail.

3.1 Sentiment Analysis or Emotion Detection

Nandwani and Verma describe sentiment analysis as a means of "assessing if data is positive, negative or neutral", and hence as the process retrieving a polarity value for certain data [8]. On the other hand, emotion detection is a an allocation of a specific category to a piece of data, like "fear", "anger" or "happiness". The field of emotion detection is a field that still needs a lot of development before it can match the usefulness and ubiquity of sentiment analysis, and it is much harder to develop a good algorithm for this purpose [13]. Also, our research question addresses the positivity and negativity of Scratch comments overall and does not require further categorization into specific emotions. These two pieces of information led to the decision to use sentiment analysis to answer the research question.

SentiStrength is a piece of software that suits our needs of automatically retrieving a value for the sentiment of a piece of text. It has human accuracy for short social texts, which is exactly the type of texts we have on a social network like Scratch [15; 14]. In fact, SentiStrength gives a score of -1 to -5 to express the negative sentiment strength and a score of 1 to 5 to express the positive sentiment strength. It gives two values, because research has shown that humans process positive and negative sentiments in parallel [1]. We will call those values the positive sentiment score and the negative sentiment score; we might also refer to them as the positive/negative sentiment or the positive/negative score.

According to Graßl and Fraser [3], SentiStrength gives a performance similar to VADER [4] and both SentiStrength and VADER performed significantly better than Standford's CoreNLP [6]. We chose to use SentiStrength over VADER because that makes it easier to build forward on the dataset of Van der Ende [16], who also used SentiStrength.

3.2 Data Collection

For the data explorations we performed, we used two datasets. The first dataset that we used is the dataset from Van der Ende [16], which we call the Ende dataset. The second dataset we used is self-generated. It can be considered as an enriched subset of the Ende dataset. We call this dataset the Schaap dataset. We describe these datasets and the generation of the Schaap dataset below.

Ende Dataset

The Ende dataset is described extensively in the paper by Van der Ende [16]. This dataset is used for data explorations 1 and 4. This is because these explorations especially looked into the effects of comment sentiments in a specific time period. The Ende dataset fits perfectly for this purpose, because it contains only projects in a specific time period, that is, the year 2019.

The dataset can be considered as an enriched subset of users and projects from the dataset from Zeevaarders and Aivaloglou [20], which can be considered a random sample due to its enormous size. As mentioned, this subset of users and projects generated by Van der Ende contains only projects that were created in 2019. The dataset contains information about 199552 Scratch projects, 772289 comments and

Exploration and dataset	Statistical Test	Independent variable	Dependent variable
Exploration 1 Ende	Pearson	Relative or absolute number of comments that have a positive or negative score (above a threshold) in timeframe 1	Number of projects created per month in timeframe 2 over the number of projects created per month in timeframe 1
Exploration 2 Schaap	Pearson	Relative or absolute number of comments that have a positive or negative score (above a threshold)	Active time, from the first project creation date to the last project modification date
Exploration 3 Schaap	Pearson	Ratio of the number of positive, negative or neutral comments over the total number of comments in active time	Active time, from the first project creation date to the last project modification date
Exploration 4 Ende	Pearson	Ratio of the number of positive, negative or neutral comments over the total number of comments in timeframes 1 and 2 together	Number of projects created in timeframes 1 and 2 together
Exploration 5 Schaap	Pearson, Spearman and Kendall Tau	Average number of comments per project	Number of projects created
Exploration 6 Schaap	Pearson, Spearman and Kendall Tau	Ratio of the number of positive, negative or neutral comments over the total number of comments in the first three projects	Number of projects coming afterwards
Exploration 7 Schaap	Pearson, Spearman and Kendall Tau	Negative or positive score for each comment	Number of projects created afterward

Table 1: Data Explorations

707669 replies. It is enriched because it contains sentiment scores for each comment.

We have decided not to use the information about the replies, because of a notion from Fields et al.: They did not explicitly mention in their research whether they scraped replies as well or whether they categorized replies on comments as comments as well. However, in their comment categories "Building a Following" and "Conversational Partners", they have included some examples of messages that seem to be replies. The sentiment of these messages seems to be related to the original comment and not to the project itself [2]. Because of this, we have decided not to take the replies to comments into account for this research.

Schaap Dataset

The Schaap dataset is used for data explorations 2, 3, 5, 6 and 7. These data explorations are about the whole period of time where users are active. This dataset fits perfectly for this purpose, because it contains information about the complete time that specific users are active on the platform.

To build the Schaap dataset, we used the Ende dataset to retrieve all the usernames belonging to the projects in that dataset. After we retrieved those usernames, we retrieved all projects (not only those from 2019) from all these users via the API [2].

After this, we removed all active accounts from the data we gathered. We used the definition of active accounts from Zeevaarders and Aivaloglou, which is based on the 95th percentile of the difference in days between subsequent project modifications. They calculated this number to be 41 [20]. With that number, we decided on a cut-off date. If there was a project modification after that cut-off date, we considered the user as active and ignored that user for the purpose of data exploration. We only considered accounts with the latest modification dates before that cut-off date; we call those accounts inactive accounts.

The code that we have written to acquire the data for this dataset can be found on GitHub [3].

The fact that the Ende dataset was generated to only include projects from 2019 does not remove the randomness aspect from the dataset of Zeevaarders and Aivaloglou. This implies that we can consider the Ende dataset as randomly collected as well, which is important when we build forward on this dataset. This fact even give us an advantage: The dataset does not contain accounts that have been inactive for many years; hence, our research data is not very old.

From the 195552 projects in the Ende dataset, we extracted 8929 users. (Later, we discovered that this number actually had to be 13970 users, but this does not influence the rest of our research, since we took a subset of this set of users later.) From these users, 13 users seem to have deleted their accounts, since no information on them can be found anymore on the Scratch platform. This left us with 8916 users. From these users, we scraped information about all their cre-

[2] https://api.scratch.mit.edu/users/⟨u⟩/projects

[3] https://github.com/gjschaap0x/cse3000

Ende dataset	Schaap dataset
Contains 199552 projects	Contains 200814 projects
Subset from Zeevaarders and Aivaloglou, enriched with sentiment data	Subset from Ende, enriched with projects that are not created in 2019
Contains only projects created in 2019	Contains, for each user, all projects that he has created

Table 2: A comparison between the Ende dataset and the Schaap dataset

ated projects, this left us with 1360871 projects. The scraping of these projects took place on November 28 and 29 in 2023.

In contrast to the generation of the Ende dataset, the project information on the Scratch website does not include the number of comments anymore [11]. This increased the time to scrape all comments on all those projects significantly, since projects that have no comments could not be skipped anymore. Therefore, we made the decision to take a subset of these 1360871 projects with about the same size as the Ende dataset. We did this in the following manner: We selected iteratively one of the users in the Ende dataset randomly (excluding users that were already selected) and we added all of their projects to the Schaap dataset, until we had selected over 200000 projects. This left us with a total of 1270 users and their 200814 projects.

From these projects, we scraped 1157350 comments via the API [4]. During scraping, 5 comments got scraped twice accidentally. This left us with a total of 1157345 comments. We performed sentiment analysis on these comments using SentiStrength, as described in section 3.1.

The Schaap dataset now contains tables similar to the Ende dataset described by Van der Ende [16], with the following exceptions:

1. The Schaap dataset does not contain a 'Reply' table and a 'ReplySentiment' table, because we did not take the replies into account, because of the reason described in the paragraphs on the generation of the Ende dataset.
2. The Schaap dataset does not contain a field in the 'Projects' table called 'comments', because it could not be scraped, as described above.
3. The Schaap dataset does not contain a field in the 'Comments' table called 'language', since Van der Ende described that it had no added value for performing sentiment analysis [16].

An overview of the two datasets is given in table 2.

3.3 Variable Selection

Dependent Variables

As stated in the research question, the dependent variables need to relate to a creator's inclination to continue creating projects. The first variable to think about would then be the number of projects created by a user, possibly normalized by the number of projects he created before receiving specific comments.

[4]https://api.scratch.mit.edu/users/⟨u⟩/projects/⟨p⟩/comments

Another variable relating to a creator's inclination to continue creating projects might be the total time that a user is active on the Scratch platform.

Data explorations were conducted with these dependent variables in mind.

Independent Variables

As stated in the research question, the independent variables need to relate to the sentiment of Scratch comments. When using the sentiment scores from SentiStrength (explained in section 3.1), we could measure the overall negativity or positivity of all comments that a user receives over a period of time either in a relative manner (e.g. the average positive sentiment is 2.1) or in an absolute manner (e.g. the user received 6 comments with a negative score of -3 or lower).

Another way to look at the sentiment of comments is using categories. Comments can be placed into different categories (positive, negative or neutral) based on their sentiment scores.

Data explorations were conducted with these independent variables in mind.

3.4 Approach

To answer the research question, we performed seven different data explorations with different variables, each aiming to investigate the potential relationship between the sentiment of comments on Scratch projects and the creator's inclination to continue producing projects. Table 1 summarizes these explorations in words and table 3 summarizes them in mathematical terms.

Expl. 1: For all the mathematical formulas below, u represents a user and is therefore in the set of all usernames in the dataset. $p(u) \in \{T(u), 1, 2, 3\}$ where $T(u)$ represents the whole time a user is active on the platform. Furthermore, $t \in \{2, 3, 4, 5\}$.

$n_{\#}(u, p(u), t)$ is the total amount of reactions that user u receives per month in time period $p(u)$ with a negative sentiment score that is lower than or equal to $-t$.

$p_{\#}(u, p(u), t)$ is the total amount of reactions that user u receives per month in time period $p(u)$ with a positive sentiment score that is larger than or equal to t.

$n_{\%}(u, p(u))$ is the average negative sentiment score of all comments that user u receives during time period $p(u)$.

$p_{\%}(u, p(u))$ is the average positive sentiment score of all comments that user u receives during time period $p(u)$.

For the first data exploration, we took the period from January 2019 to (including) May 2019. We did not extend the period to August, because statistics on the Scratch website show different activity trends for the summer months compared to months during the school year [12]. To bypass this effect, the last month that we included for the period is May 2019. We divided this period into two periods. The first period is from January 2019 to (including) April 2019, while the second period is May 2019. We hypothesized that users who receive a lot of positive comments create more projects as a result of those comments. We wanted to see whether there was any relationship between the comment sentiments of the comments that users would receive in the first period with the number of projects they created per month in the second period.

$N(u, p(u))$ is the average number of projects created per month by user u during time period $p(u)$.

This means that, for the first data exploration, we compared $n_{\#}(u, 1, t)$, $p_{\#}(u, 1, t)$, $n_{\%}(u, 1)$ and $p_{\%}(u, 1)$ with $\frac{N(u,2)}{N(u,1)}$ for all possible values of t.

Expl. 2: For the second data exploration, we created the Schaap dataset that contains, for a randomly selected group of users, information about all their projects and all the comments they have received on their projects. We hypothesized that users who receive many positive comments would be longer active on the platform. We define the active time as the time between the first project creation date and the last project modification date of a user.

$T(u)$ is the time between the first project creation date of user u and his last project edit date. In other words, it is the active time of a user.

This means that, for the second data exploration, we compared $n_{\#}(u, T(u), t)$, $p_{\#}(u, T(u), t)$, $n_{\%}(u, T(u))$ and $p_{\%}(u, T(u))$ with $T(u)$ for all possible values of t.

Expl. 3: After the first and second data exploration, we found some drawbacks of the measurements that were used regarding the overall negativity or positivity of comments. Instead of measuring overall negativity or positivity relatively or absolutely, we decided to label each comment as being "negative", "positive" or "neutral". Then, we looked at the ratio of comments in each category and compared that with the active time of the user.

A comment is defined as positive if and only if it has a positive score of 3 or higher. A comment is defined as negative if and only if it is not defined as positive and it has a negative score of -3 or lower. A comment is defined as neutral (unbiased) if and only if it is neither defined as positive nor as negative.

$n_c(u, p(u))$ is the ratio of the number of negative comments that user u receives during time period $p(u)$ to the total number of comments that user u receives in this time period.

$p_c(u, p(u))$ is the ratio of the number of positive comments that user u receives during time period $p(u)$ to the total number of comments that user u receives in this time period.

$u_c(u, p(u))$ is the ratio of the number of neutral (unbiased) comments that user u receives during time period $p(u)$ to the total number of comments that user u receives in this time period.

This means that, for the third data exploration, we compared $p_c(u, T(u))$, $n_c(u, T(u))$ and $u_c(u, T(u))$ with $T(u)$.

Expl. 4: In the first data exploration, it turned out that the timeframes used were too short. Instead of looking at shorter timeframes, we decided to take a look at a longer timeframe. Here, we used the Ende dataset again. We defined the third timeframe as the combination of the first and second timeframe as defined in the first exploration, which is therefore the period from January 2019 until (including) May 2019. We decided to continue using the categorical approach from the third data exploration.

This means that, for the fourth data exploration, we compared $p_c(u, 3)$, $n_c(u, 3)$ and $u_c(u, 3)$ with $N(u, 3)$.

Expl. 5: We hypothesized that users who receive more comments create on average more projects than other users (regardless of their sentiment).

$C(u)$ is the average number of comments that user u receives per project.

$P(u)$ is the number of projects that user u has created in total.

This means that, for the fifth data exploration, we compared $C(u)$ with $P(u)$.

Expl. 6: We hypothesized that the comments that a user receives on his first projects have more influence on the number of projects that he creates compared to comments that a user receives later. We decided to check this for the first three projects of each user.

Like the categorical variables, $n_3(u, p(u))$, $p_3(u, p(u))$ and $u_3(u, p(u))$ are the ratio of the number of negative, positive and neutral comments that user u receives during time period $p(u)$ to the total number of comments that user u receives in this time period, but only on the first 3 projects that user u creates in that time period.

This means that, for the sixth data exploration, we compared $p_3(u, T(u))$, $n_3(u, T(u))$ and $u_3(u, T(u))$ with $P(u) - 3$.

Expl. 7: We decided to look at the research question from a different perspective: Instead of checking for every user what the effect of the comments is, we wanted to check for every comment what the effect on the number of projects created afterwards is.

$p(c) \in \{1, 2, 3, 4, 5\}$ is the positive sentiment score of comment c.

$n(c) \in \{-1, -2, -3, -4, -5\}$ is the negative sentiment score of comment c.

$P(c)$ is the number of projects that a project author creates after receiving comment c.

This means that, for the seventh data exploration, we compared $p(c)$ and $n(c)$ with $P(c)$.

For data explorations 1, 2, 3 and 4 we generated Pearson correlation coefficients between the independent and the dependent variables. For data explorations 5, 6 and 7, we also generated Spearman Rank correlation coefficients and Kendall Tau correlation coefficients. In all data explorations, the null hypothesis was that there does not exist a correlation and the alternative hypothesis was that there does exist a correlation.

4 Results

The results of all data explorations are given in table 4. Note that all results that indicate a correlation are significant on a two-tailed T-test with significance level $\alpha < 0.05$.

For some independent variables, a graph is shown in this report. A selection of graphs has been made that represents the results well. Also, some graphs are limited in the horizontal and vertical direction (and hence do not contain a few outliers), with the purpose of showing most of the data points well. In the graphs, $T(u)$ is represented in days.

For the first data exploration, a correlation for the variables could not be found. When looking at the graph for $p_{\#}(u, 1, 2)$ in figure 1, we suspected that our original hypothesis, that how more positive comments a user receives, how more projects he creates, was not true. In this graph,

	Ind. var.	Dep. var.
Expl. 1	$n_{\#}(u,1,t)$ $p_{\#}(u,1,t)$ $n_{\%}(u,1)$ $p_{\%}(u,1)$	$\frac{N(u,2)}{N(u,1)}$
Expl. 2	$n_{\#}(u,T(u),t)$ $p_{\#}(u,T(u),t)$ $n_{\%}(u,T(u))$ $p_{\%}(u,T(u))$	$T(u)$
Expl. 3	$p_c(u,T(u))$ $n_c(u,T(u))$ $u_c(u,T(u))$	$T(u)$
Expl. 4	$p_c(u,3)$ $n_c(u,3)$ $u_c(u,3)$	$N(u,3)$
Expl. 5	$C(u)$	$P(u)$
Expl. 6	$p_3(u,T(u))$ $n_3(u,T(u))$ $u_3(u,T(u))$	$P(u)-3$
Expl. 7	$p(c)$ $n(c)$	$P(c)$

Table 3: Data Explorations in Mathematical Definitions

users who receive fewer negative comments seem to create more projects than other users. It might even be true that users who create lots of comments take longer to finish their projects, and hence create less projects. This made us realize that a number of projects created might not really be a good measurement for measuring a creator's inclination to continue producing projects.

For the variables in the second data exploration, a very low correlation was found, with the exception of $n_{\%}(u,T(u))$. However, as can be seen in figure 2, this correlation is not clearly visible on the graphs.

Also, we were surprised by the fact that the median active time was more than five years, as can be seen in table 6. When exploring this remarkable phenomenon, we found many users that have been inactive for a long period of time and then got active again. An example of these users is shown in figure 3.

For the variables in the third data exploration, a very low correlation for the variables could be found. However, as can be seen in figure 4, this correlation is not clearly visible in the graphs.

For the fourth data exploration, a correlation could not be found, as can be seen in figure 5. Although it was remarkable to see that there were many data points at specific values of $p_c(u,3)$, $n_c(u,3)$ and $u_c(u,3)$, namely 0.0, 1.0, 0.5, 0.75, 0.25 and some others. These "spikes" were because of users that received a very small number of comments from January until (including) May 2019, which makes the ratio of the number of comments of a certain sentiment to the total number of comments drop into one of these "spikes".

For the fifth data exploration, a low correlation could be found, as can be seen in figure 6. Our conjecture from the first data exploration, about the inverse relationship between the number of comments that a user receives and the number of projects that a user received was consolidated by this ex-

	Ind. var.	Results	Meaning
Expl. 1	$n_{\#}(u,1,2)$	P: -0.00	No corr.
	$n_{\#}(u,1,3)$	P: -0.01	No corr.
	$n_{\#}(u,1,4)$	P: -0.01	No corr.
	$n_{\#}(u,1,5)$	P: -0.01	No corr.
	$p_{\#}(u,1,2)$	P: -0.00	No corr.
	$p_{\#}(u,1,3)$	P: -0.01	No corr.
	$p_{\#}(u,1,4)$	P: -0.01	No corr.
	$p_{\#}(u,1,5)$	P: -0.01	No corr.
	$n_{\%}(u,1)$	P: 0.01	No corr.
	$p_{\%}(u,1)$	P: 0.02	No corr.
Expl. 2	$n_{\#}(u,T(u),2)$	P: -0.11	Low corr.
	$n_{\#}(u,T(u),3)$	P: -0.11	Low corr.
	$n_{\#}(u,T(u),4)$	P: -0.11	Low corr.
	$n_{\#}(u,T(u),5)$	P: -0.11	Low corr.
	$p_{\#}(u,T(u),2)$	P: 0.12	Low corr.
	$p_{\#}(u,T(u),3)$	P: 0.12	Low corr.
	$p_{\#}(u,T(u),4)$	P: 0.12	Low corr.
	$p_{\#}(u,T(u),5)$	P: 0.12	Low corr.
	$n_{\%}(u,T(u))$	P: -0.01	No corr.
	$p_{\%}(u,T(u))$	P: -0.14	Low corr.
Expl. 3	$p_c(u,T(u))$	P: -0.17	Low corr.
	$n_c(u,T(u))$	P: -0.11	Low corr.
	$u_c(u,T(u))$	P: 0.19	Low corr.
Expl. 4	$p_c(u,3)$	P: -0.04	No corr.
	$n_c(u,3)$	P: 0.01	No corr.
	$u_c(u,3)$	P: 0.04	No corr.
Expl. 5	$C(u)$	P: -0.05	No corr.
		S: -0.18	Low corr.
		K: -0.13	Low corr.
Expl. 6	$p_3(u,T(u))$	P: -0.06	No corr.
		S: -0.09	No corr.
		K: -0.06	No corr.
	$n_3(u,T(u))$	P: 0.02	No corr.
		S: -0.05	No corr.
		K: -0.04	No corr.
	$u_3(u,T(u))$	P: 0.06	No corr.
		S: 0.07	No corr.
		K: 0.04	No corr.
Expl. 7	$p(c)$	P: -0.05	No corr.
		S: 0.03	No corr.
		K: 0.02	No corr.
	$n(c)$	P: 0.00	No corr.
		S: -0.01	No corr.
		K: -0.01	No corr.

Table 4: Results of Data Explorations. In the Results column, P stands for Pearson correlation coefficient, S stands for Spearman rank correlation coefficient and K stands for Kendall Tau correlation coefficient. All the results indicating a correlation are significant on a two-tailed T-test with significance level $\alpha < 0.05$.

Min.	0.00
Q1	0.00
Q2	0.50
Q3	1.16
Max.	48.00

Table 5: Five-number summary of $\frac{N(u,2)}{N(u,1)}$ for Expl. 1

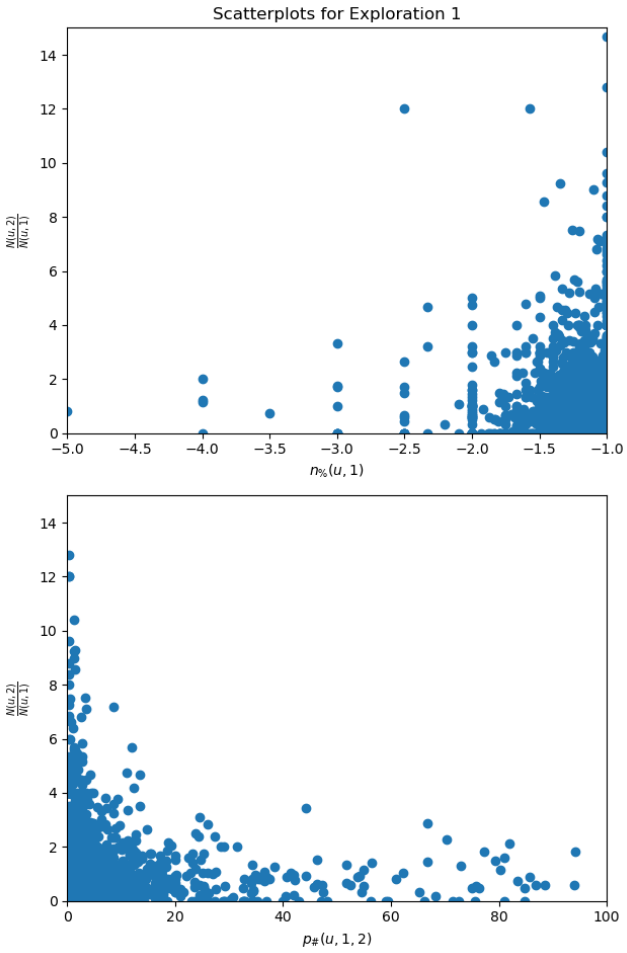

Figure 1: Scatterplots for Exploration 1

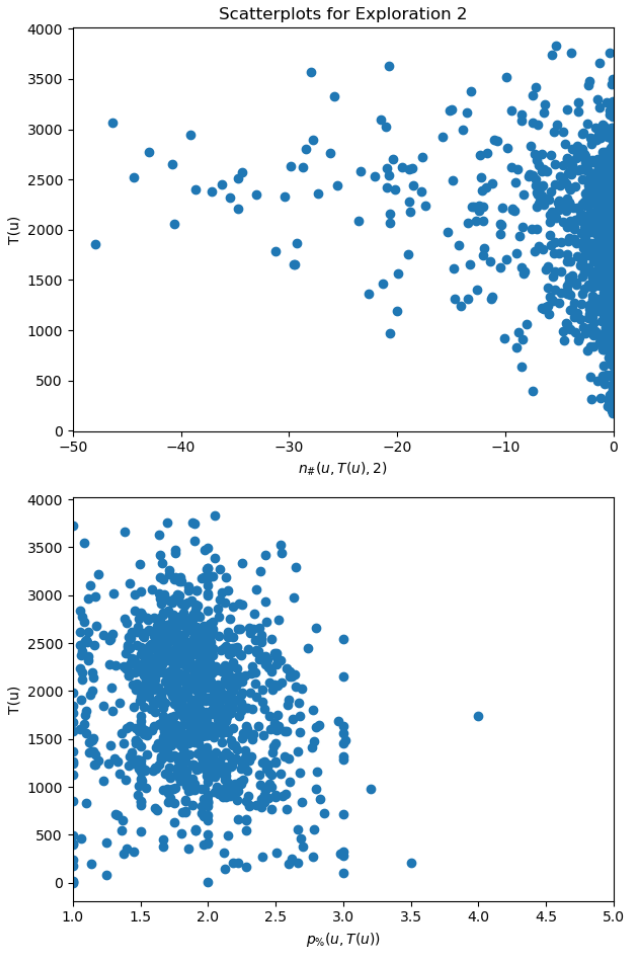

Figure 2: Scatterplots for Exploration 2

ploration. Interestingly, only the Spearman and Kendal-Tau correlation coefficients revealed a correlation, while the Pearson correlation coefficient did not, as can be seen in table 4.

For the sixth data exploration, no correlation could be found, as can be seen in figure 7.

For the seventh data exploration, no correlation could be found, as can be seen in figure 8. Interestingly, the median values and even the third quartile values for the positive and negative scores of each comment are low compared to the maximum values found. The distributions contain a large amount of outliers, as can be seen when comparing the scatterplots on the left and the boxplots on the right of figure 8.

In short, a significant low correlation could be found for data explorations 2, 3 and 5. No correlation could be found for data exploration 1, 4, 6 and 7.

Min.	0
Q1	1313.5
Q2	1909
Q3	2405
Max.	3828

Table 6: Five-number summary of $T(u)$ for Expl. 2 and 3

5 Responsible Research

In section 5.1, we discuss the ethical considerations regarding the data that was used for this research. In section 5.2, we discuss the reproducibility of this research.

5.1 Data Usage

The data used for this research comes from either the Ende dataset or the Schaap dataset. Both datasets are constructed through the process of scraping the Scratch website. The only data that is scraped is data that is publicly available.

According to the privacy policy of Scratch, users are asked "not to share personal contact information in projects, comments, profiles, studios or forum posts" [10]. This leaves the responsibility of leaving personal information at a publicly available place to the user. Nevertheless, the Privacy Policy explains that users could use the "Report" button if they encounter personal information. Also, it explains that users in the EEA, UK and Switzerland have other rights to officially request a restriction of processing personal data.

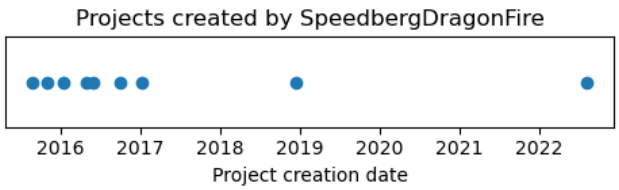

Figure 3: Example of a user who has been inactive for a long time

It can be questioned whether children (who are the vast majority of Scratch users) understand these privacy risks when they are active on the Scratch platform. Due to the size of the Ende dataset and the Schaap dataset, it can also not be guaranteed that some users nevertheless left personal identifiable information in some comments or project titles. When the Scratch platform itself receives a request to restrict the processing of the personal data of some user, the research team will not receive this request, since the datasets are already scraped and generated. This holds for both comments and projects when they are reported for containing personal information.

For these reasons, it is decided to not make the Schaap dataset publicly available online. When another researcher requests personally to access the Schaap dataset, it will be given to him, personally, to ensure confidentiality.

5.2 Reproducibility

This study is designed in a reproducible manner. The data acquisition process is described extensively. Furthermore, all code used to process this data for the different data explorations will be made publicly available.

In addition, the contents of the Schaap dataset can be retrieved on request, as explained in section 5.1.

6 Discussion

As described in section 4, some explorations gave no correlation, other explorations gave a low correlation. In this section, we discuss several limitations on the research done. In section 6.1, we discuss the definition of active time we used. In section 6.2, we discuss the reliability of some data explorations. In section 6.3, we mention the possible influence of a "Report" button on this research. In section 6.4, we mention some other factors that could have influenced this research.

6.1 Active Time

As mentioned in section 3.2, for the creation of the Schaap dataset we used the definition of active time from Zeevaarders and Aivaloglou [20]: They made a distribution of the time between the creation of two projects and they decided on a cut-off date at the 95th percentile of this distribution, which was 41 days. All users that have created a project in the last 41 days before the data was scraped were considered active users, while the others were considered inactive users. During this research, we have run into disadvantages of using this definition, which were revealed after discovering many users who were inactive for a long time and then got active again on the platform. The activity of one of these users is shown

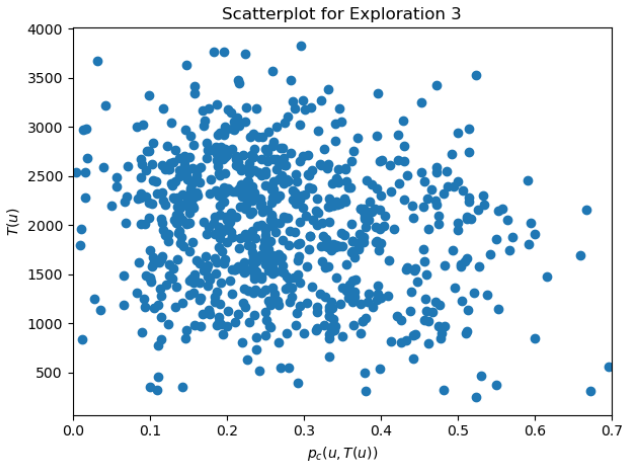

Figure 4: Scatterplot for Exploration 3

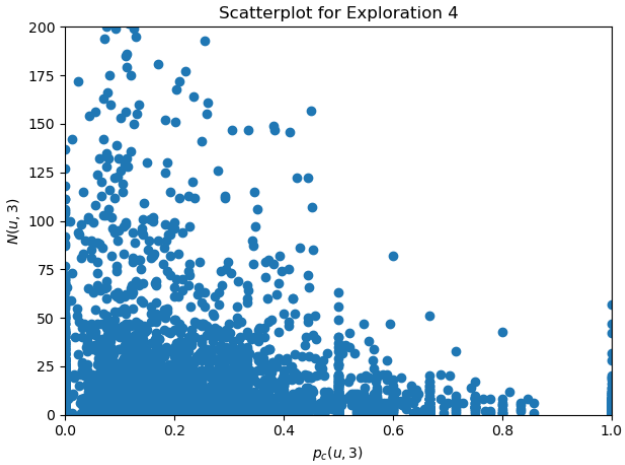

Figure 5: Scatterplot for Exploration 4

in figure 3. That user is technically considered inactive, since the last time he created a project was more than 41 days ago, but the question arises whether he really is inactive, since he recently created a project again after more than three years of not creating any project.

The Zeevaarders and Aivaloglou definition of active time places more weight on users who create many projects, compared to users who create fewer projects. We think that this has significantly reduced the number at the 95th percentile. For example, if there is one user who creates one thousand projects per year and there are ten users who create three projects per year, those ten users could all easily be considered inactive, even though they are active.

A solution to this disadvantage is to take a look at the average time between the creation of two consecutive projects for each user individually, and then take a look at the distribution of all those averages for each user, and then take the 95th percentile out of that distribution.

Another solution could be to create a distribution for each

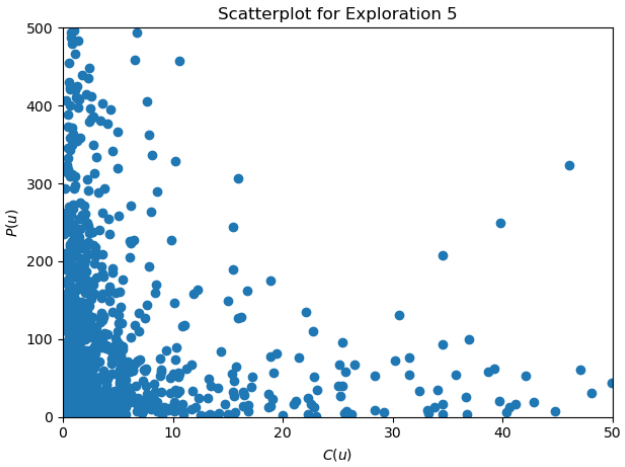

Figure 6: Scatterplot for Exploration 5

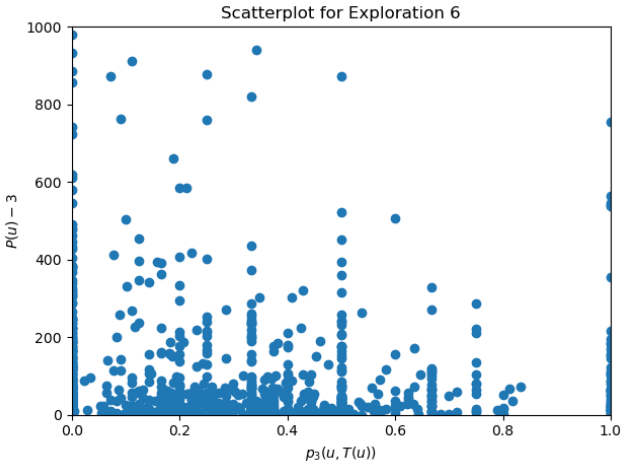

Figure 7: Scatterplot for Exploration 6

user and to take the 95th percentile for each user individually to define whether he is inactive.

During this research, it was not possible to try these solutions due to time constraints. It can be considered as a recommendation for future work.

6.2 Reliability of explorations

Regarding first data exploration, as can be seen in table 5, the third quartile of the spread of $\frac{N(u,2)}{N(u,1)}$ is just above 1.0, while the second quartile is 0.5. This means that there might be a large number of people who did not even create one project in May 2019 and therefore their average number of projects created per month is below 1. This might have influenced our results for data exploration 1 heavily and this makes these results unreliable.

Regarding the sixth data exploration, even much stronger than in the fourth data exploration, "spikes" are visible throughout the graph in figure 7. This might be because many children do not receive many comments on their very first

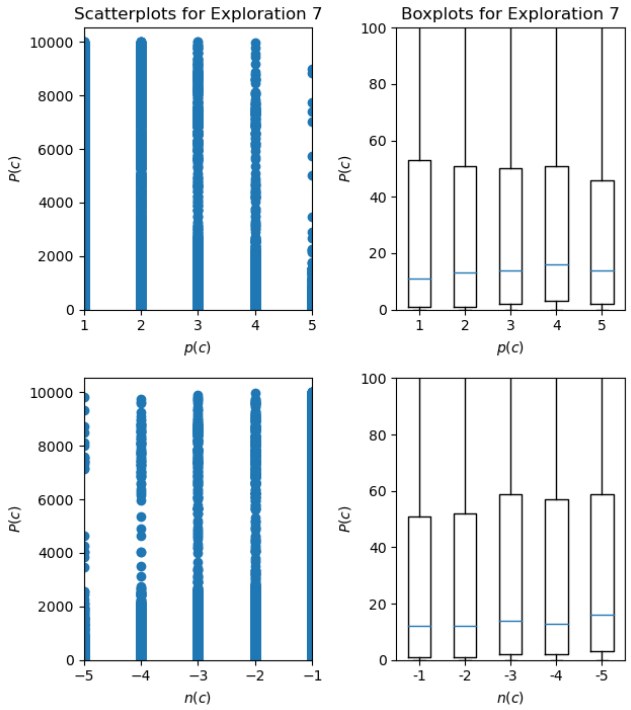

Figure 8: Scatterplots and Boxplots for Exploration 7

projects, especially not if they are just starting to program and their skills are not yet developed. Because of the appearance of many "spikes", we will consider these results as unreliable.

6.3 Reporting of Negative Comments

Comments on the Scratch platform have a "Report" button. It might be the case that the most negative or the most discouraging comments (including mean, insulting and offensive comments) are already reported, and therefore they could not have been taken into account for this research.

For example, user "doodlebug5" has created a project called "GAY is OKAY." In the comments section, there are multiple people mentioning that they see some people are being rude in the comments section [5]. However, these rude comments can not be found back in the comments section. The fact that many rude comments have already been reported might have impacted our results.

6.4 Other confouding factors

Some things that might have influenced the results of this research as well are the following:

- Originally, the Schaap dataset originates from the Ende dataset and the Ende dataset originates from the dataset from Zeevaarders and Aivaloglou. The dataset from Zeevaarders and Aivaloglou is scraped recursively, starting with a random set of users and then continuing to scrape friends. Because they have scraped friends of friends, it might be that there are not many hated projects

[5] https://scratch.mit.edu/projects/177712104/

in the dataset. It could be true that a user does not want to follow some other user if he does not like or even hates his projects.

- It might be the case that children make more spelling mistakes than adults. Also, it might be the case that they use more unofficial language that is not (yet) in a dictionary, such as slang. This might have influenced the sentiment scores calculated by SentiStrength, because it might not have recognized the words correctly. This reduces the confidence we have on the sentiment scores that were calculated.

7 Conclusions

This research has done data analysis to explore the relation between the sentiment of comments on a Scratch project and the creator's inclination to continue producing projects. In seven different iterations, it has explored the relationship between different variables relating to the sentiment of comments and to a creator's inclination to continue producing projects. This research has shown that the effect of sentiment alone on the creator's inclination to continue producing projects is very limited, no correlation was found in 4 out of 7 iterations and a low correlation was found in 3 out of 7 iterations.

For future research regarding the time that a user is active on the platform, this research recommends to try another way of calculating whether or not users are still active on the platform, because the way from Zeevaarders and Aivaloglou turns out to have its drawbacks.

Furthermore, the fact that most of the rude comments on Scratch are already reported made this research less reliable. Also, the way of scraping projects (scraping friends of friends) might have excluded the most hated projects from users who have almost no friends on the Scratch platform. Next to this, the fact that children might make more spelling mistakes than adults might have influenced the sentiment scores calculated by SentiStrength.

References

[1] Raul Berrios, Peter Totterdell, and Stephen Kellett. Eliciting mixed emotions: A meta-analysis comparing models, types, and measures. *Frontiers in psychology*, 6:428, 2015.

[2] Deborah A Fields, Katarina Pantic, and Yasmin B Kafai. "i have a tutorial for this" the language of online peer support in the scratch programming community. In *Proceedings of the 14th International Conference on Interaction Design and Children*, pages 229–238, 2015.

[3] Isabella Graßl and Gordon Fraser. Scratch as social network: topic modeling and sentiment analysis in scratch projects. In *Proceedings of the 2022 ACM/IEEE 44th International Conference on Software Engineering: Software Engineering in Society*, pages 143–148, 2022.

[4] C Hutto and E Gilbert. A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8:216–225, 2014.

[5] Mizuko Ito, Heather A Horst, Matteo Bittanti, Becky Herr Stephenson, Patricia G Lange, CJ Pascoe, Laura Robinson, et al. *Living and learning with new media: Summary of findings from the digital youth project*. The MIT Press, 2009.

[6] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[7] Anthony Martinez and Cheridan Christnacht. Women are nearly half of us workforce but only 27% of stem workers. *United States Census Bureau*, 2021.

[8] Pansy Nandwani and Rupali Verma. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):81, 2021.

[9] Gabriela T Richard and Yasmin B Kafai. Blind spots in youth diy programming: Examining diversity in creators, content, and comments within the scratch online community. In *Proceedings of the 2016 CHI conference on Human Factors in Computing Systems*, pages 1473–1485, 2016.

[10] Scratch. Privacy policy. Retrieved 17 January 2024 at https://scratch.mit.edu/privacy_policy.

[11] Scratch. Scratch api documentation. Retrieved 7 December 2023 at https://en.scratch-wiki.info/wiki/Scratch_API/.

[12] Scratch. Scratch statistics. Retrieved 17 November 2023 at https://scratch.mit.edu/statistics/.

[13] Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*, 2018.

[14] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.

[15] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*, 61(12):2544–2558, 2010.

[16] Dyon van der Ende. A dataset of comments and their sentiment. Bachelor's thesis, Leiden University, 2020.

[17] Nicole Forsgren Velasquez, Deborah A Fields, David Olsen, Taylor Martin, Mark C Shepherd, Anna Strommer, and Yasmin B Kafai. Novice programmers talking about projects: What automated text analysis reveals about online scratch users' comments. In *2014 47th Hawaii international conference on system sciences*, pages 1635–1644. IEEE, 2014.

[18] Fu-Hsiang Wen, Tienhua Wu, and Wei-Chih Hsu. Toward improving student motivation and performance in introductory programming learning by scratch: The role of achievement emotions. *Science Progress*, 106(4):00368504231205985, 2023.

[19] Zhanxia Yang and Marina Bers. Examining gender difference in the use of scratchjr in a programming curriculum for first graders. *Computer Science Education*, pages 1–22, 2023.

[20] Ad Zeevaarders and Efthimia Aivaloglou. Exploring the programming concepts practiced by scratch users: an analysis of project repositories. In *2021 IEEE Global Engineering Education Conference (EDUCON)*, pages 1287–1295. IEEE, 2021.