Alice Pucher, Bsc.

StackExchange: The effectiveness of the new contributor indicator, an analysis

Master's Thesis

to achieve the university degree of

Master of Science

Master's degree programme: Computer Science

submitted to

Graz University of Technology

Supervisor

Assoc.Prof. Dipl.-Ing. Dr.techn. Denis Helic

Institute for Interactive Systems and Data Science
Head: Univ.-Prof. Dipl.-Inf. Dr. Stefanie Lindstaedt

Graz, August 2020

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

Date	Signature

Abstract

StackExchange is a question and answer platform and like other social platforms, StackExchange is eager to provide a good first impression to users. StackExchange made many decisions to attract new users. One of these decisions was to introduce the *new contributor* indicator which is shown to users that may answer a question from a new user. This thesis investigates whether this change improved the impression, new users experience. To measure whether the change achieved its intended target, this thesis uses VADER to quantify the sentiment of the answers to questions of new contributors which are then used in an interrupted time series. The results indicate that in some of the communities the change did indeed achieve its intended purpose.

Contents

List of Figures

1 Introduction

StackExchange is a Q&A platform and consists of 174 communities[1]. Each community evolves around a specific topic, for instance, StackOverflow focusing on software engeneering, or AskUbuntu focusing on the Ubuntu operating system. This distincts StackExchange from other Q&A sites such as *Yahoo! Answers* where no such differentiation into topics exists.

In August of 2018, the StackExchange team introduced a small change that may have had a huge impact on the platform. They added a new feature to visibly mark questions from new contributors, as part of their effort to make the site more welcoming for new users[2]. Specifically, members who want to answer a question created by a new contributor are shown a notification in the answer box that this question is from a new contributor. The StackExchange team hopes that this little change encourages members to be more friendly and forgiving toward new users.

This thesis evaluates whether this change has a real impact on the community and if so how the community reacts. For this analysis, this thesis utilizes Vader [1], a sentiment analysis tool, to measure the sentiments of the answers submitted to questions of new contributors. Furthermore, this thesis includes the votes these questions receive and the number of questions new contributors ask. Interrupted time series are then applied to these values to evaluate whether the change achieved its purpose of making the platform more welcoming.

This thesis investigates the ten largest communities of the StackExchange platform measured by the number of posts. This includes most prominent communities, for instance, StackOverflow, MathOverflow, Math, AskUbuntu, and SuperUser as well as some lesser-known communities.

[1] https://stackexchange.com/tour

[2] https://meta.stackexchange.com/questions/314287/come-take-a-look-at-our-new-contributor-indicator

The remaining part of this thesis is structured as follows: Section 2 explains Stack-Exchange and its communities, how it works, and shows related work. Section 3 shows the method this thesis uses for analysis in detail. Section 4 contains the investigated datasets. Results are presented in Section 5 and discussed in Section 6. Section 7 concludes this thesis.

2 Related Work

This section is divided into three parts. The first part explains what StackExchange is, how it developed since its inception, and how it works. The second part shows previous and related work. The third section covers approaches to analyze sentiment as well as methods to analyze trends over time.

2.1 Background

StackExchange[1] is a community question and answering (CQA) platform where users can ask and answer questions, accept answers as an appropriate solution to the question, and up-/downvote questions and answers. StackExchange uses a community-driven knowledge creation process by allowing everyone who registers to participate in the community. Invested users also get access to moderation tools to help maintain the vast community. All posts on the StackExchange platform are publicly visible, allowing non-users to benefit from the community as well. Posts are also accessible for web search engines so users can find questions and answers easily with a simple web search. StackExchange keeps an archive of all questions and answers posted, creating a knowledge archive for future visitors to look into.

Originally, StackExchange started with StackOverflow[2] in 2008[3]. Since then StackExchange grew into a platform hosting sites for 174 different topics[4], for instance, programming (StackOverflow), maths (MathOverflow[5] and Math StackExchange[6]),

[1] https://stackexchange.com
[2] https://stackoverflow.com
[3] https://stackoverflow.blog/2008/08/01/stack-overflow-private-beta-begins/
[4] https://stackexchange.com/tour
[5] https://mathoverflow.net
[6] https://math.stackexchange.com

and typesetting (TeX/LaTeX[7]). Questions on StackExchange are stated in natural English language and consist of a title, a body containing a detailed description of the problem or information need, and tags to categorize the question. After a question is posted the community can submit answers to the question. The author of the question can then accept an appropriate answer which satisfies their question. The accepted answer is then marked as such with a green checkmark and shown on top of all the other answers. Figure 2.1 shows an example of a StackOverflow question. Questions and answers can be up-/downvoted by every user registered on the site. Votes typically reflect the quality and importance of the respective question or answers. Answers with a high voting score raise to the top of the answer list as answers are sorted by the vote score in descending order by default. Voting also influences a user's reputation [2][4]. When a post (question or answers) is voted upon the reputation of the poster changes accordingly. Furthermore, downvoting of answers also decreases the reputation of the user who voted[8].

Reputation on StackExchange indicates how trustworthy a user is. To gain a high reputation value a user has to invest a lot of time and effort to reach a high reputation value by asking good questions and posting good answers to questions. Reputation also unlocks privileges which may differ slightly from one community to another[9,10]. With privileges, users can, for instance, create new tags if the need for a new tag arises, cast votes on closing or reopening questions if the question is off-topic or a duplicate of another question, or when a question had been closed for no or a wrong reason, or even get access to moderation tools. StackExchange also employs a badge system to steer the community[11]. Some badges can be obtained by performing one-time actions, for instance, reading the tour page which contains necessary details for newly registered users, or by performing certain actions multiple times, for instance, editing and answering the same question within 12 hours. Furthermore, users can comment on every question and answer. Comments could be used for further clarifying an answer or a short discussion on a question or answer.

For each community on StackExchange, a *M*eta page is offered where members of the respective community can discuss the associated community [3][12]. This place is

[7] https://tex.stackexchange.com
[8] https://stackoverflow.com/help/privileges/vote-down
[9] https://mathoverflow.com/help/privileges/
[10] https://stackoverflow.com/help/privileges/
[11] https://stackoverflow.com/help/badges/
[12] https://stackoverflow.com/help/whats-meta/

used by site admins to interact with the community. The *M*eta pages are also used for proposing and voting on new features and reporting bugs. *M*eta pages run the same software as the normal CQA pages so users vote on ideas and suggestions in the same way they would do on the actual CQA sites.

2.2 State of the Art

Since the introduction of Web 2.0 and the subsequential spawning of platforms for social interaction, researchers started investigating emerging online communities. Research strongly focuses on the interactions of users on various platforms. Community knowledge platforms are of special interest, for instance, StackExchange/StackOverflow [4, 5, 6, 2, 7, 8, 9, 10, 11, 12], Quora [13], Reddit [14, 15], Yahoo! Answers [16, 17], and Wikipedia [18]. These platforms allow communication over large distances and facilitate fast and easy knowledge exchange and acquisition by connecting thousands or even millions of users and creating valuable repositories of knowledge in the process. Users create, edit, and consume little pieces of information and collectively build a community and knowledge repository. However, not every piece of information is factual [13, 16] and platforms often employ some kind of moderation to keep up the value of the platform and to ensure a certain standard within the community.

All these communities differ in their design. Wikipedia is a community-driven knowledge repository and consists of a collection of articles. Every user can create an article. Articles are edited collaboratively and continually improved and expanded. Reddit is a platform for social interaction where users create posts and comment on other posts or comments. Quora, StackExchange, and Yahoo! Answers are community question and answer (CQA) platforms. On Quora and Yahoo! Answers users can ask any question regarding any topics whereas on StackExchange users have to post their questions in the appropriate subcommunity, for instance, StackOverflow for programming-related questions or MathOverflow for math-related questions.

CQA sites are very effective at code review [19]. Code may be understood in the traditional sense of source code in programming-related fields but this also translates to other fields, for instance, mathematics where formulas represent code. CQA sites are also very effective at solving conceptual questions. This is due to the

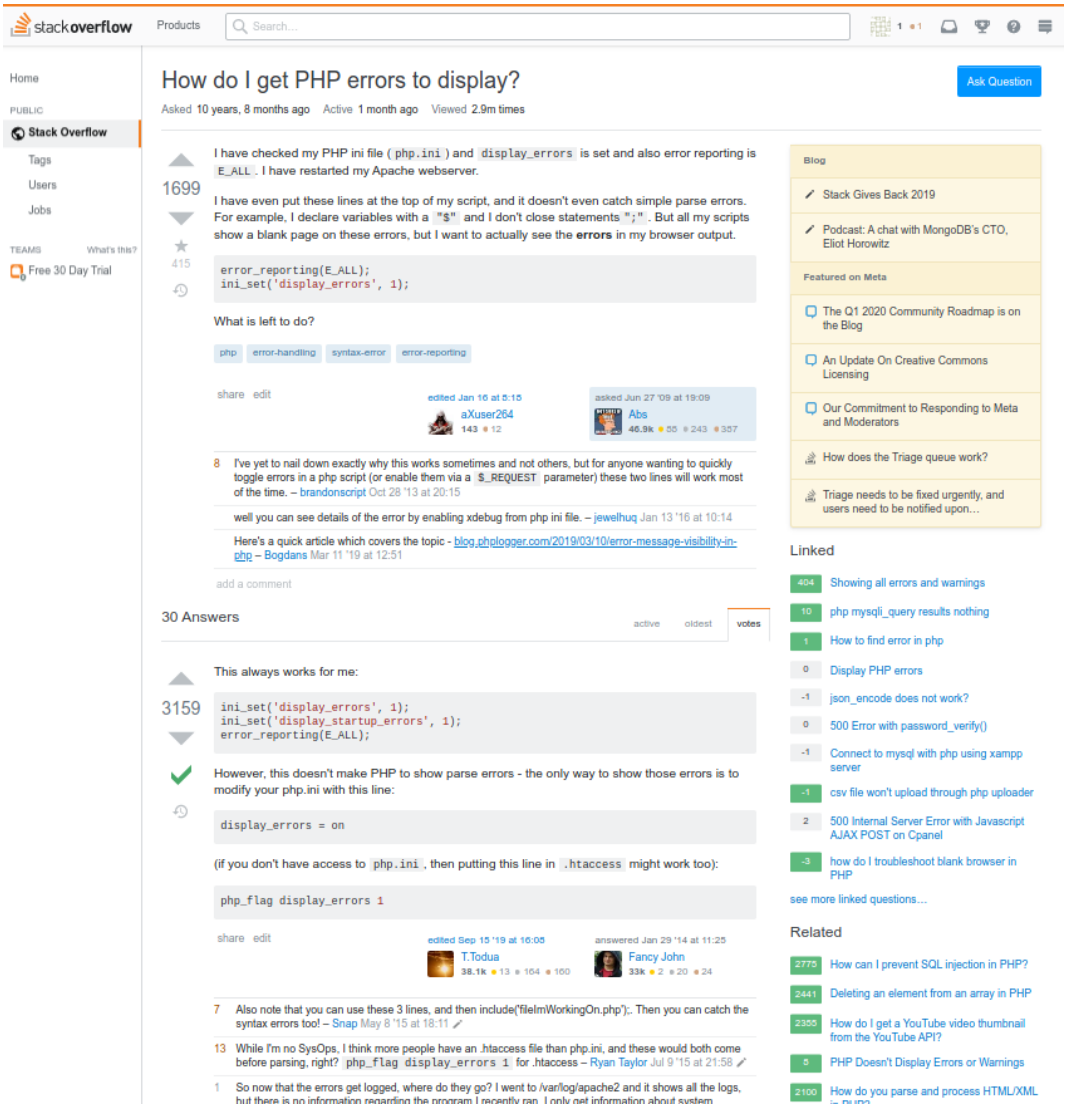


Figure 2.1: A typical question on StackOverflow. In the top middle section of the page, the question is stated. The question has 4 tags and 3 comments attached to it. Beneath the question, all answers are listed by their score in descending order (only one answer is visible in this screenshot). The accepted answer is marked by a green checkmark. To the left of the question and answers, the score (computed via votes) is indicated.

fact that people have different areas of knowledge and expertise [20] and due to the large user base established CQA sites have, which again increases the variety of users with expertise in different fields.

2.2.1 Running an online community

Despite the differences in purpose and manifestation of these communities, they are social communities and they have to follow certain laws. In their book on "Building successful online communities: Evidence-based social design" [21] Kraut and Resnick lie out five equally important criteria online platforms have to fulfill in order to thrive:

1) When starting a community, it has to have a critical mass of users who create content. StackOverflow already had a critical mass of users from the beginning due to the StackOverflow team already being experts in the domain [3] and the private beta[3]. Both aspects ensured a strong community core early on.

2) The platform must attract new users to grow as well as replace leaving users. Depending on the type of community new users should bring certain skills, for example, programming background in open-source software development, or extended knowledge on certain domains; or qualities, for example, a certain illness in medical communities. New users also bring the challenge of onboarding with them. Most newcomers will not be familiar with all the rules and nuances of the community [18][13].

3) The platform should encourage users to commit to the community. Online communities are often based on the voluntary commitment of their users [22], hence the platform has to ensure users are willing to stay. Most platforms do not have contracts with their users, so users should see benefits for staying with the community.

4) Contribution by users to the community should be encouraged. Content generation and engagement are the backbones of an online community.

5) The community needs regulation to sustain it. Not every user in a community is interested in the well-being of the community. Therefore, every community has to

[13]https://stackoverflow.blog/2018/04/26/stack-overflow-isnt-very-welcoming-its-time-for-that-to-change/

deal with trolls and inappropriate or even destructive behavior. Rules need to be established and enforced to limit and mitigate the damage malicious users cause.

All these criteria are heavily intertwined. Attracting new users often depends on the welcomingness and support of users that are already on the platform. Keeping users committed to the platform depends on the engagement with the community and how well the system design supports this. The following sections cover the criteria 2) to 5).

2.2.2 Onboarding

The onboarding process of new users is a permanent challenge for online communities and differs from one platform to another. New users should be welcomed by the community and helped to integrate themselves into the community. This is a continuous process. It is not enough for a user to make one contribution and then revert to a non-contributing state. The StackExchange team took efforts to onboard new users better by making several changes to the site. However, there are still problems where further actions are required.

One-day-flies
Slag, Waard and Bacchelli investigated why many users on StackOverflow only post once after their registration [4]. They found that 47% of all users on StackOverflow posted only once and called them one-day-flies. They suggest that code example quality is lower than that of more involved users, which often leads to answers and comments to first improve the question and code instead of answering the stated question. This likely discourages new users from using the site further. Negative feedback instead of constructive feedback is another cause for discontinuation of usage. The StackOverflow staff also conducted their own research on negative feedback of the community[14]. They investigated the comment sections of questions by recruiting their staff members to rate a set of comments and they found more than 7% of the reviewed comments are unwelcoming.

One-day-flies are not unique to StackOverflow. Steinmacher et al. investigated the social barriers newcomers face when they submit their first contribution to an open-source software project [23]. They based their work on empirical data and interviews and identified several social barriers preventing newcomers to place

[14]https://stackoverflow.blog/2018/07/10/welcome-wagon-classifying-comments-on-stack-overflow/

their first contribution to a project. Furthermore, newcomers are often on their own in open source projects. The lack of support and peers to ask for help hinders them. Yazdanian et al. found that new contributors on Wikipedia face challenges when editing articles. Wikipedia hosts millions of articles[15] and new contributors often do not know which articles they could edit and improve. Recommender systems can solve this problem by suggesting articles to edit but they suffer from the cold start problem because they rely on past user activity which is missing for new contributors. Yazdanian et al. proposed a solution by establishing a framework that automatically creates questionnaires to fill this gap. This also helps match new contributors with more experienced contributors that could help newcomers when they face a problem. Allen showed that the one-time-contributors phenomenon also translates to workplaces and organizations [24]. They found out that socialization with other members of an organization plays an important role in turnover. The better the socialization within the organization the less likely newcomers are to leave. This socialization process has to be actively pursued by the organization.

Lurking
One-day-flies may partially be a result of lurking. Lurking is consuming content generated by a community but not contributing content to it. Nonnecke, Andrews and Preece investigated lurking behavior on Microsoft Network (MSN) [25] and found that contrary to previous studies [26, 27] lurking is not necessarily a bad behavior. Lurkers show passive behavior and are more introverted and less optimistic than actively posting members of a community. Previous studies suggested lurking is free riding, a taking-rather-than-giving process. However, the authors found that lurking is important in getting to know a community, how a community works, and learning the nuances of social interactions on the platform. This allows for better integration into the community when a person decides to join the community. StackExchange, and especially the StackOverflow community, probably has a large lurking audience. Many programmers do not register on the site and those who do only ask one question and revert to lurking, as suggested by [4].

Reflection
The StackOverflow team acknowledged the one-time-contributors trend[13,14] and took efforts to make the site more welcoming to new users[16]. They lied out various reasons: Firstly, they have sent mixed messages whether the site is an expert site or

[15]https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia
[16]https://stackoverflow.blog/2018/06/21/rolling-out-the-welcome-wagon-june-update/

for everyone. Secondly, they gave too little guidance to new users which resulted in poor questions from new users and in the unwelcoming behavior of more integrated users towards the new users. New users do not know all the rules and nuances of communication of the communities. An example is that ”Please“ and ”Thank you“ are not well received on the site as they are deemed unnecessary. Also the quality, clearness, and language quality of the questions of new users is lower than more experienced users which leads to unwelcoming or even toxic answers and comments. Moreover, users who gained moderation tool access could close questions with predefined reasons which often are not meaningful enough for the poster of the question[17]. Thirdly, marginalized groups, for instance, women and people of color [28][13,18], are more likely to drop out of the community due to unwelcoming behavior from other users[13]. They feel the site is an elitist and hostile place. The team suggested several steps to mitigate these problems. Some of these steps include appealing to the users to be more welcoming and forgiving towards new users[13,14,19], other steps are geared towards changes to the platform itself: The *Be nice policy* (code of conduct) was updated with feedback from the community[20]. This includes: new users should not be judged for not knowing all things. Furthermore, the closing reasons were updated to be more meaningful to the poster, and questions that are closed are shown as ”on hold“ instead of ”closed“ for the first 5 days[17]. Moreover, the team investigates how the comment sections can be improved to lessen the unwelcomeness and hostility and keep the civility up.

Mentorship Research Project
The StackOverflow team partnered with Ford et al. and implemented the Mentorship Research Project [5][21]. The project lasted one month and aimed to help newcomers improve their first questions before they are posted publicly. The program went as follows: When a user is about to post a question the user is asked whether they want their question to be reviewed by a mentor. If they confirmed they are forward to a help room with a mentor who is an experienced user. The question is then reviewed and the mentor suggests some changes if applicable. These changes may include narrowing the question for more precise answers, adding a code

[17]https://stackoverflow.blog/2013/06/25/the-war-of-the-closes/
[18]https://insights.stackoverflow.com/survey/2019
[19]https://stackoverflow.blog/2012/07/20/kicking-off-the-summer-of-love/
[20]https://meta.stackexchange.com/questions/240839/the-new-new-be-nice-policy-code-of-conduct-updated-wit
[21]https://meta.stackoverflow.com/questions/357198/mentorship-research-project-results-wrap-up

example or adjusting code, or removing of *Please* and *Thank you* from the question. After the review and editing, the question is posted publicly by the user. The authors found that mentored questions are received significantly better by the community than non-mentored questions. The questions also received higher scores and were less likely to be off-topic and poor in quality. Furthermore, newcomers are more comfortable when their question is reviewed by a mentor. For this project, four mentors were hand-selected and therefore the project would not scale very well as the number of mentors is very limited but it gave the authors an idea on how to pursue their goal of increasing the welcomingness on StackExchange. The project is followed up by a *Ask a question wizard* to help new users, as well as more experienced users, improve the structure, quality, and clearness of their questions[16].

Unwelcomeness
Unwelcomeness is a large problem on StackExchange [28][16,13]. Although unwelcomeness affects all new users, users from marginalized groups suffer significantly more [29][13]. Ford et al. investigated barriers users face when contributing to StackOverflow. The authors identified 14 barriers in total hindering newcomers to contribute and five barriers were rated significantly more problematic for women than men. On StackOverflow only 5.8% (2015[22], 7.9% 2019[18]) of active users identify as women. David and Shapiro found similar results of 5% women in their work on *Community-based production of open-source software* [30]. These numbers are comparatively small to the number of degrees in Science, Technology, Engineering, and Mathematics (STEM) [31] where 20% are achieved by women [32]. Despite the difference, the percentage of women on StackOverflow has increased in recent years.

2.2.3 Invoke commitment

While attracting and onboarding new users is an important step for growing a community, keeping them on the platform and turning them into long-lasting community members is equally as important for growth as well as sustainability. Users have to feel the benefits of staying with the community. Without the benefits, a user has little to no motivation to interact with the community and will most likely drop out of it. Benefits are diverse, however, they can be grouped into 5 categories:

[22]https://insights.stackoverflow.com/survey/2015

information exchange, social support, social interaction, time and location flexibility, and permanency [33].

As StackExchange is a CQA platform, the benefits from information exchange, time and location flexibility, and permanency are more prevalent, while social support and social interaction are more in the background. Social support and social interaction are more relevant in communities where individuals communicate about topics regarding themselves, for instance, communities where health aspects are the main focus [34]. Time and location flexibility is important for all online communities. Information exchange and permanency are important for StackExchange as it is a large collection of knowledge that mostly does not change over time or from one individual to another. StackExchange's content is driven by the community and therefore depends on the voluntarism of its users, making benefits even more important.

The backbone of a community is always the user base and its voluntarism to participate with the community. Even if the community is led by a commercial core team, the community is almost always several orders of magnitude greater than the number of the paid employees forming the core team [35]. The core team often provides the infrastructure for the community and does some community work. However, most of the community work is done by volunteers of the community. This is also true for the StackExchange platform where the core team of paid employees is between 200 to 500[23] (this includes employees working on other products) and the number of voluntary community members (these users have access to moderation tools) performing community work is around 10,000 [24].

2.2.4 Encourage contribution

In a community, users can generally be split into 2 groups by motivation to voluntarily contribute: One group acts out of altruism, where users contribute with the reason to help others and do good to the community; the second group acts out of egoism and selfish reasons, for instance, getting recognition from other people [36]. Users of the second group still help the community but their primary

[23] https://www.linkedin.com/company/stack-overflow

[24] https://data.stackexchange.com/stackoverflow/revision/1412005/1735651/users-with-rep-20k

goal is not necessarily the health of the community but gaining reputation and making a name for themselves. Contrary, users of the first group primarily focus on helping the community and see reputation as a positive side effect which also feeds back in their ability to help others. While these groups have different objectives, both groups need recognition of their efforts [33]. There are several methods for recognizing the value a member provides to the community: reputation, awards, trust, identity, etc. [36]. Reputation, trust, and identity are often reached gradually over time by continuously working on them, awards are reached at discrete points in time. Awards often take some time and effort to achieve. However, awards should not be easily achievable as their value comes from the work that is required for them[37]. They should also be meaningful in the community they are used in. Most importantly, awards have to be visible to the public, so other members can see them. In this way, awards become a powerful motivator to users.

StackExchange employs several features to engage users with the platform, for instance, the reputation system and the badge (award) system. These systems reward contributing users with achievements and encourage further contribution to the community. Both systems try to keep and increase the quality of the posts on the platform.

Reputation
Reputation plays an important role on StackExchange and indicates the credibility of a user, as well as a primary source of answers of high-quality [2]. Although the largest chunk of all questions is posted by low-reputation users, high-reputation users post more questions on average. To earn a high reputation a user has to invest a lot of effort and time into the community, for instance, asking good questions or providing useful answers to questions of others. Reputation is earned when a question or answer is upvoted by other users, or if an answer is accepted as the solution to a question by the question creator. Mamykina et al. found that the reputation system of StackOverflow encourages users to compete productively [3]. But not every user participates equally, and participation depends on the personality of the user [6]. Bazelli, Hindle and Stroulia showed that the top-reputation users on StackOverflow are more extroverted compared to users with less reputation. Movshovitz-Attias et al. found that by analyzing the StackOverflow community network, experts can be reliably identified by their contribution within the first few months after their registration. Graph analysis also allowed the authors to find spamming users or users with other extreme behavior.

Although gaining reputation takes time and effort, users can take certain advantages to gain reputation faster by gaming the system [7, 38]. Bosu et al. analyzed the reputation system and found five strategies to increase the reputation in a fast way: Firstly, answering questions with tags that have a small expertise density. This reduces competitiveness against other users and increases the chance of upvotes and answer acceptance. Secondly, questions should be answered promptly. The question asker will most likely accept the first arriving answer that solves the question. This is also supported by [39]. Thirdly, answering first also gives the user an advantage over other answerers. Fourthly, activity during off-peak hours reduces the competition from other users. Finally, contributing to diverse areas will also help in developing a higher reputation. This behavior may, however, decrease answer quality when users focus too much on reputation collection and disregard the quality of their posts[38].

Badges
Complementary to the reputation system StackOverflow also employs a badge system[11] to stimulate contributions by users [40]. The goal of badges is to keep users engaged with the community [41]. Therefore, badges are often used in a gamification setting where users contribute to the community and are rewarded for their behavior if it aligns with the requirements of the badges. Badges are visible in questions and answers as well as the profile page of the user and can be earned by performing certain actions. Badges are often seen as a steering mechanism by researchers [8, 9, 10]. Although users want to achieve badges and are therefore steered to perform certain actions, steering also occurs in the reputation system. However, badges allow a wider variety of goals, for instance, asking and answering questions, voting on questions and answers, or writing higher-quality answers.

Badges also work as a motivator for users [10]. Users often put in non-trivial amounts of work and effort to achieve badges and so badges become powerful incentives. However, not all users are equal and therefore do not pursue badges in the same way [8]. Contrary to [10], Yanovsky et al. [8] found that users do not necessarily increase their activity prior to achieving a badge followed by an immediate decrease in contribution thereafter but users behave differently based on their type of contribution. The authors found users can be categorized into three groups: Firstly, some users are not affected at all by the badge system and still contribute a lot to the community. Secondly, users increase their activity too before gaining a badge and keep their level of contribution afterward. Finally, users

increase their activity before achieving a badge and return to their previous level of engagement thereafter.

Different badges also create status classes [11]. The harder a badge can be earned by users, the more unique it is within the community and therefore the badge symbolizes some sort of status. Often rare badges are hard to achieve and take significant effort. For some users, depending on their type, this can be a huge motivator. Kusmierczyk and Gomez-Rodriguez found first-time badges play an important role in steering users [9]. The steering effect only takes place if the benefit to the user is greater than the effort the user has to put in to obtain the badge. If the effort is greater the user will likely not pursue the badge and therefore the steering effect will not occur.

2.2.5 Regulation

Regulation evolves around the user actions and the content a community creates. It is required to steer the community and keep the community civil. Naturally, some users will not have the best intentions for the community in mind. These actions of such must be accounted for, and harmful actions must be dealt with. Otherwise, the community and its content will deteriorate.

Content quality
Quality is a concern in online communities. Platform moderators and admins want to keep a certain level of quality or even raise it. However, higher-quality posts take more time and effort than lower-quality posts. In the case of CQA platforms, this is an even bigger problem as higher-quality answers fight against fast responses. Despite that, StackOverflow also has a problem with low quality and effort questions and the subsequent unwelcoming answers and comments[14].

Lin et al. investigated how growth affects a community[14]. They looked at Reddit communities that were added to the default set of subscribed communities of every new user (defaulting) which lead to a huge influx of new users to these communities as a result. The authors found that contrary to expectations, the quality stays largely the same. The vote score dips shortly after defaulting but quickly recovers or even raises to higher levels than before. The complaints of low-quality content did not increase, and the language used in the community stayed the same. However, the community clustered around fewer posts than before defaulting. Srba and Bielikova

did a similar study on the StackOverflow community [38]. They found a similar pattern in the quality of posts. The quality of questions dipped momentarily due to the huge influx of new users. However, the quality did recover after 3 months.

Tausczik and Pennebaker found reputation is linked to the perceived quality of posts in multiple ways [12]. They suggest reputation could be used as an indicator of quality. Quality also depends on the type of platform. Lin et al. showed that expert sites who charge fees, for instance, library reference services, have higher quality answers compared to free sites[14]. Also, the higher the fee the higher the quality of the answers. However, free community sites outperform expert sites in terms of answer density and responsiveness.

Content abuse
Srba and Bielikova identified 3 types of users causing the lowering of quality: *Help Vampires* (these spend little to no effort to research their questions, which leads to many duplicates), *Noobs* (they create mostly trivial questions), and *Reputation Collectors*[38]. They try to gain reputation as fast as possible by methods described by Bosu et al.[7] but often with no regard of what effects their behavior has on the community, for instance, lowering overall content quality, turning other users away from the platform, and encouraging the behavior of *Help Vampires* and *Noobs* even more.

Questions of *Help Vampires* and *Noobs* direct answerers away from much more demanding questions. On one hand, this leads to knowledgeable answerers answering questions for which they are overqualified to answer, and on the other hand to a lack of adequate quality answers for more difficult questions. Srba and Bielikova suggest a system that tries to match questions with answerers that satisfy the knowledge requirement but are not grossly overqualified to answer the question. A system with this quality would prevent suggesting simple questions to overqualified answerers, and prevent an answer vacuum for questions with more advanced topics. This would ensure more optimal utilization of the answering capability of the community.

Content moderation
Srba and Bielikova proposed some solutions to improve the quality problems. One suggestion is to restrict the openness of a community. This can be accomplished in different ways, for instance, introducing a posting limit for questions on a daily basis[38]. While this certainly limits the amount of low-quality posts, it does not eliminate the problem. Furthermore, this limitation would also hurt engaged users

which would create a large volume of higher quality content. A much more intricate solution that adapts to user behavior would be required, otherwise, the limitation would hurt the community more than it improves.

Ponzanelli et al. performed a study where they looked at post quality on StackOverflow[42]. They aim to improve the automatic low-quality post detection system which is already in place and reduce the size of the review queue selected individuals have to go through. Their classifier improves by including popularity metrics of the user posting and the readability of the post itself. With these additional factors, they managed to reduce the amount of misclassified quality posts with only a minimal decrease of correctly classified low-quality posts. Their improvement to the classifier reduced the review queue size by 9%.

Another solution is to find content abusers (noobs, help vampires, etc.) directly. One approach is to add a reporting system to the community, however, a system of this kind is also driven by user inputs and therefore can be manipulated as well. This would lead to excluding users flagged as false positives and missing a portion of content abusers completely. A better approach is to systematically find these users by their behavior. Kayes et al. describe a classifier which achieves an accuracy of 83% on the *Yahoo! Answers* platform [17]. The classifier is based on empirical data where they looked at historical user activity, report data, and which users were banned from the platform. From these statistics, they created the classifier which is able to distinguish between falsely and fairly banned users. Cheng, Danescu-Niculescu-Mizil and Leskovec performed a similar study on antisocial behavior on various platforms. They too looked at historical data of users and their eventual bans as well as on their deleted posts rates. Their classifier achieved an accuracy of 80%.

2.3 Analysis

When analyzing a community, one typically finds 2 types of data: text, and metadata. Metadata is relatively easy to quantify, while text is much more complicated and intricate to quantify. Text contains a large variety of features and depending on the research in question, researchers have to decide which features they want to include. This thesis investigates the (un-)friendliness in the communication between users and will therefore perform sentiment analysis on the texts. The next section will

go into more detail on sentiment analysis. After the data (text and metadata) is quantified, one often wants to know how the data has changed over time. The trend analysis section follows the sentiment analysis section.

2.3.1 Sentiment analysis

Researchers put forth many tools for sentiment analysis over the years. Each tool has its advantages and drawbacks and there is not a silver bullet solution that fits all research questions. Researchers have to choose a tool that best fits their needs and they need to be aware of the drawbacks of their choice. Sentiment analysis poses three important challenges:

- Coverage: detecting as many features as possible from a given piece of text
- Weighting: assigning one or multiple values (value range and granularity) to detected features
- Creation: creating and maintaining a sentiment analysis tool is a time and labor-intensive process

In general, sentiment analysis tools can be grouped into two categories: handcrafted and automated (machine learning).

Handcrafted Approches
Creating hand-crafted tools is often a huge undertaking. They depend on a hand-crafted lexicon (gold standard, human-curated lexicons), which maps features of a text to a value. In the simplest sense, these just map a word to a binary value -1 (negative word) or 1 (positive word). However, most tools use a more complex lexicon to capture more features of a piece of text. By design, they allow a fast computation of the sentiment of a given piece of text. Also, hand-crafted lexicons are easy to update and extend. Furthermore, hand-crafted tools produce easily comprehensible results. The following paragraphs explain some of the analysis tools in this category.

Linguistic Inquiry and Word Count (LIWC) [44, 45] is one of the more popular tools. Due to its widespread usage, LIWC is well verified, both internally and externally. Its lexicon consists of about 6,400 words where words are categorized into one or more of the 76 defined categories [46]. 620 words have a positive and 744 words have a negative emotion. Examples for positive words are: love, nice, sweet; examples for negative words are: hurt, ugly, nasty. LIWC also has some drawbacks,

for instance, it does not capture acronyms, emoticons, or slang words. Furthermore, LIWC's lexicon uses a polarity-based approach, meaning that it cannot distinguish between the sentences "This pizza is good" and "This pizza is excellent"[1]. *G*ood and *e*xcellent are both in the category of positive emotion but LIWC does not distinguish between single words in the same category.

General Inquirer (GI)[47] is one of the oldest sentiment tools still in use. It was originally designed in 1966 and has been continuously refined and now consists of about 11000 words where 1900 positively rated words and 2300 negatively rated words. Like LIWC, GI uses a polarity-based lexicon and therefore is not able to capture sentiment intensity[1]. Also, GI does not recognize lexical features, such as acronyms, initialisms, etc.

Hu-Liu04 [48, 49] is a opinion mining tool. It searches for features in multiple pieces of text, for instance, product reviews, and rates the opinion of the feature by using a binary classification[48]. Crucially Hu-Liu04 does not summarize the texts but summarizes ratings of the opinions about features mentioned in the texts. Hu-Liu04 was bootstrapped from WordNet[48] and then extended further. It now uses a lexicon consisting of about 6800 words where 2000 words have a positive sentiment and 4800 words have a negative sentiment attached[1]. This tool is, by design, better suited for social media texts, although it also misses emoticons, acronyms, and initialisms.

SenticNet [50] is also an opinion mining tool but it focuses on concept-level opinions. SenticNet is based on a paradigm called *Sentic Mining* which uses a combination of concepts from artificial intelligence and the Semantic Web. More specifically, it uses graph mining and dimensionality reduction. SenticNets lexicon consists of about 14250 common-sense concepts which have ratings on many scales of which one is a polarity score with a continuous range from -1 to 1[1]. This continuous range of polarity scores enables SenticNet to be sentiment-intensity aware.

Affective Norms for English Words (ANEW) [51] is a sentiment analysis tool and was introduced to standardize research and offer a way to compare research. Its lexicon is fairly small and consists of only 1034 words which are ranked pleasure, arousal, and dominance. However, ANEW uses a continuous scale from 1 to 9 where 1 represents the negative end, 9 represents the positive end, and 5 is considered neutral. With this design, ANEW is able to capture sentiment intensity. However, ANEW still misses lexical features, for instance, acronyms[1].

WordNet analyzes text with a dictionary that contains lexical concepts [52, 53]. Each lexical concept contains multiple words which are synonyms, called synsets. These synsets are then linked by semantic relations. With this lexicon, text can be queried in multiple different ways.

SentiWordNet [54] is an extension of WordNet and adds sentiment scores to the synsets. Its lexicon consists of about 147000 synsets, each having 3 values (positive, neutral, negative) attached to them. Each value has a continuous range from 0 to 1 and the sum of these 3 values is set to be 1. The values of each synset are calculated by a mix of semi-supervised algorithms, mostly propagation, and classifiers. This distinguishes SentiWordNet from previously explained sentiment tools, where the lexica are exclusively created by humans (except for simple mathematical operations, for instance, averaging of values). Therefore, SentiWordNet's lexicon is not considered to be a human-curated gold standard. Furthermore, the lexicon is very noisy and most of the synsets are neither positive nor negative but a mix of both[1]. Moreover, SentiWordNet misses lexical features, for instance, acronyms, initialisms, and emoticons.

Word-Sense Disambiguation (WSD)[55] is not a sentiment analysis tool per se, however, it can be used to enhance others. In languages certain words have different meanings depending on the context they are used in. When sentiment tools, which do not use WSD, analyze a piece of text, some words which have different meanings depending on the context may skew the resulting sentiment. Some words can even change from positive to negative or vice versa depending on the context. WSD tries to distinguish between subjective and objective word usage. For example *The party was great.* and *The party lost many votes.* Although *p*arty is written exactly the same it has 2 completely different meanings. Depending on the context, ambiguous words can have different sentiments.

Machine Learning Approches
Because handcrafting sentiment analysis requires a lot of effort, researchers turned to approaches that offload the labor-intensive part to machine learning (ML). However, this results in a new challenge, namely: gathering a *g*ood data set to feed the machine learning algorithms for training. Firstly, *g*ood data set needs to represent as many features as possible, otherwise, the algorithm will not recognize it. Secondly, the data set has to be unbiased and representative for all the data of which the data set is a part of. The data set has to represent each feature in an appropriate amount, otherwise, the algorithms may discriminate a feature in favor of other

more represented features. These requirements are hard to fulfill and often they are not[1]. After a data set is acquired, a model has to be learned by the ML algorithm, which is, depending on the complexity of the algorithm, a very computational-intensive and memory-intensive process. After training is completed, the algorithm can predict sentiment values for new pieces of text, which it has never seen before. However, due to the nature of this approach, the results cannot be comprehended by humans easily if at all. ML approaches also suffer from a generalization problem and therefore cannot be transferred to other domains without accepting a bad performance, or updating the training data set to fit the new domain. Updating (extending or modifing) the model also requires complete retraining from scratch. These drawbacks make ML algorithms useful only in narrow situations where changes are not required and the training data is static and unbiased.

The Naive Bayes (NB) classifier is one of the simplest ML algorithms. It uses Bayesian probability to classify samples. This requires the assumption that the probabilities of the features are independent of one another, which often they are not because languages have certain structures of features.

Maximum Entropy (ME) is a more sophisticated algorithm. It uses an exponential model and logistic regression. It distinguishes itself from NB by not assuming conditional independence of features. It also supports weighting of features by using the entropy of features.

Support Vector Machines (SVM) uses a different approach. SVMs put data points in an n-dimentional space and differentiate them with hyperplanes ($n-1$ dimensional planes), so data points fall in 1 of the 2 halves of the space divided by the hyperplane. This approach is usually very memory and computation-intensive as each data point is represented by an n-dimentional vector where n denotes the number of trained features.

In general, ML approaches do not provide an improvement over hand-crafted lexicon approaches as they only shift the time-intensive process to training data set collections. Furthermore, lexicon-based approaches seem to have progressed further in terms of coverage and feature weighting. However, many tools are not specifically tailored to social media text analysis and leak in coverage of feature detection.

VADER
This shortcoming was addressed by Hutto and Gilbert who introduced a new senti-

ment analysis tool: Valence Aware Dictionary for sEntiment Reasoning (VADER)[1]. Hutto and Gilbert acknowledged the problems that many tools have and designed VADER to leverage the shortcomings. Their aim was to introduce a tool that works well in the social media domain, provides good coverage of features occurring in the social media domain (acronyms, initialisms, slang, etc.), and is able to work with online streams (live processing) of texts. VADER is also able to distinguish between different meanings of words (WSD) and it is able to take sentiment intensity into account. These properties make VADER an excellent choice when analyzing sentiment in the social media domain.

2.3.2 Trend analysis

When introducing a change to a system (experiment), one often wants to know whether the intervention achieves its intended purpose. This leads to 3 possible outcomes: a) the intervention shows an effect and the system changes in the desired way, b) the intervention shows an effect and the system changes in an undesired way, or c) the system did not react at all to the change. There are multiple ways to determine which of these outcomes occur. To analyze the behavior of the system, data from before and after the intervention as well as the nature of the intervention has to be acquired. The are multiple ways to run such an experiment and one has to choose which type of experiment fits best. There are 2 categories of approaches: actively creating an experiment where one design the experiment before it is executed (for example randomized control trials in medical fields), or using existing data of an experiment that was not designed beforehand, or where setting up a designed experiment is not possible (quasi-experiment).

As this thesis investigates a change that has already been implemented by another party, this thesis covers quasi-experiments. A tool that is often used for this purpose is an *Interrupted Time Series* (ITS) analysis. The ITS analysis is a form of segmented regression analysis, where data from before, after, and during the intervention is regressed with separate line segements[56]. ITS requires data at (regular) intervals from before and after the intervention (time series). The interrupt signifies the intervention and the time of when it occurred must be known. The intervention can be at a single point in time or it can be stretched out over a certain time span. This property must also be known to take it into account when designing the regression. Also, as the data is acquired from a quasi-experiment, it may be

baised[57], for example, seasonality, time-varying confounders (for example, a change in measuring data), variance in the number of single observations grouped together in an interval measurement, etc. These biases need to be addressed if present. Seasonality can be accounted for by subtracting the average value of each of the months in successive years (i.e. subtract the average value of all Januaries in the data set from the values in Januaries). This removes the differences between different months of the same year thereby filtering out the effect of seasonality. The variance in data density per interval (data samples in an interval) can be addressed by using each single data point in the regression instead of an average.

3 Method

StackExchange introduced a *new contributor* indicator to all communities on 21^{st} of August in 2018 at 9 pm UTC[1]. This step is one of many StackExchange took to make the platform and its members more welcoming towards new users. This indicator is shown to potential answerers in the answer text box of a question from a new contributor, as shown in figure 3.1. The indicator is added to a question if the question is the first contribution of the user or if the first contribution (question or answer) of the user was less than 7 days ago[2]. The indicator is then shown for 7 days from the creation date of the question. Note that the user can be registered for a long time and then post their first question and it is counted as a question from a new contributor. Also, if a user decides to delete all their existing contributions from the site and then creates a new question this question will have the *new contributor* indicator attached. The sole deciding factor for the indicator is the date and time of the first non-deleted contribution and the 7-day window afterward.

This thesis investigates the following criteria to determine whether the change affected a community positively or negatively, or whether the community is largely unaffected:

- **Sentiment of answers to a question**. This symbolizes the quality of communication between different individuals. Better values indicate better communication. Through the display of the *new contributor* indicator, the answerer should react less negatively towards the new user when they behave outside the community standards.
- **Vote score of questions**. This symbolizes the feedback the community gives to a question. Voters will likely vote more positively (not voting instead of down-voting, or upvoting instead of not voting) due to the *new contributor* indicator. Thereby the vote score should increase after the change.

[1] https://meta.stackexchange.com/questions/314287/come-take-a-look-at-our-new-contributor-indicator
[2] https://meta.stackexchange.com/questions/314472/what-are-the-exact-criteria-for-the-new-contributor-indicator-to-be-shown

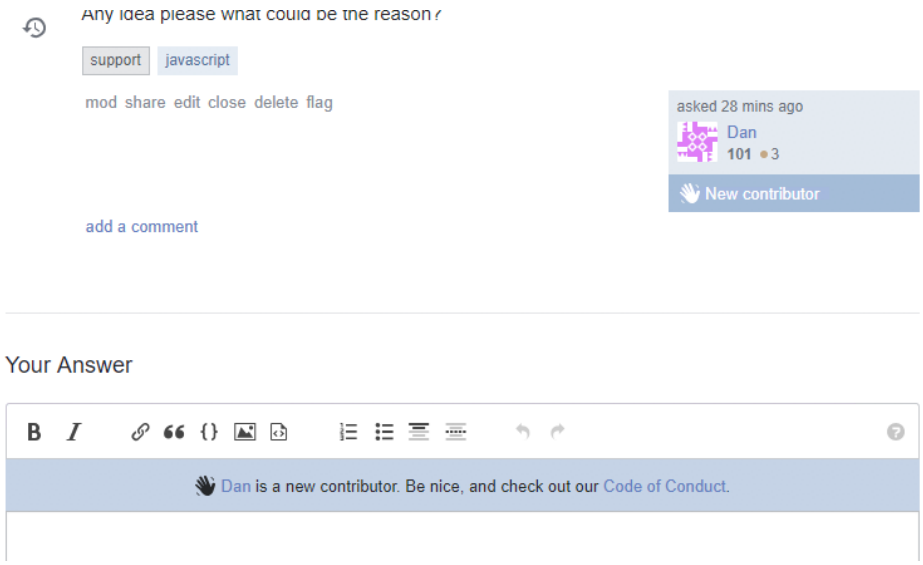


Figure 3.1: The answer box a potential answerers sees when viewing a question from a new contributor. ©Tim Post, 2018, https://meta.stackexchange.com/users/50049/tim-post[1]

- **Amount of first and follow-up question**. This symbolizes the willingness of users to participate in the community. Higher amounts of first questions indicate a higher number of new participating users. Higher follow-up questions indicate that users are more willing to stay within the community and continue their active participation.

If these criteria improve after the change is introduced, the community is affected positively. If they worsen, the community is affected negatively. If the criteria stay largely the same, then the community is unaffected. Here it is important to note that a question may receive answers and votes after the *new contributor* indicator is no longer shown and therefore these are not considered as part of the data set to analyze.

To measure the effect on the sentiment of the change this thesis utilizes the Vader[1] sentiment analysis tool. This decision is based on the performance in analyzing and categorizing microblog-like texts, the speed of processing, and the simplicity of use. Vader uses a lexicon of words, and rules related to grammar and syntax. This lexicon was manually created by Hutto and Gilbert and is therefore considered a *gold standard lexicon*. Each word has a sentiment value attached to it. Negative

words, for instance, *e*vil, have negative values; good words, for instance, *b*rave, have positive values. The range of these values is continuous, so words can have different intensities, for instance, *b*ad has a higher value than *e*vil. This feature of intensity distinction makes Vader a valance-based approach.

However, just simply looking at the words in a text is not enough and therefore Vader also uses rules to determine how words are used in conjunction with other words. Some words can boost other words. For example, "They did well." is less intense than "They did extremely well.". This works for both positive and negative sentences. Moreover, words can have different meanings depending on the context, for instance, "Fire provides warmth." and "Boss is about to fire an employee." This feature is called *Word Sense Disambiguation*.

Furthermore, Vader also detects language features commonly found in social media text which may not be present in other forms of text, for instance, books, or newspapers. Social media texts may contain acronyms, initialisms (for instance *afaik* (as far as I know)), slang words, emojis, caps words (often used to emphasize meaning), punctuation (for instance, *!!!*, and *?!?!*), etc.. These features can convey a lot of meaning and drastically change the sentiment of a text. After all these features are considered, Vader assigns a sentiment value between -1 and 1 on a continuous range. The sentiment range is divided into 3 classes: negative (-1 to -0.05), neutral (-0.05 to 0.05), and positive (0.05 to 1). The outer edges of this range are rarely reached as the text would have to be extremely negative or positive which is very unlikely.

Due to this mathematical simplicity, Vader is really fast when computing a sentiment value for a given text. This feature is one of the requirements Hutto and Gilbert originally posed. They proposed that Vader shall be fast enough to do online (real-time) analysis of social media text. Vader is also easy to use. It does not require any pre-training on a dataset as it already has a human-curated lexicon and rules related to grammar and syntax. Therefore the sentiment analysis only requires an input to evaluate. This thesis uses a publicly available implementation of Vader.[3] The design of Vader allows fast and verifiable analysis.

[3]https://github.com/cjhutto/vaderSentiment

3.1 Data gathering and preprocessing

StackExchange provides anonymized data dumps of all their communities for researchers to investigate at no cost on archive.org[4]. These data dumps contain users, posts (questions and answers), badges, comments, tags, votes, and a post history containing all versions of posts. Each entry contains the necessary information, for instance, id, creation date, title, body, and how the data is linked together (which user posted a question/answer/comment). However, not all data entries are valid and therefore cannot be used in the analysis, for instance, questions or answers of which the user is unknown, but this only affects a very small amount of entries. So before the actual analysis, the data has to be cleaned. Moreover, the answer texts are in HTML format, containing tags that could skew the sentiment values, and they need to be stripped away beforehand. Additionally, answers may contain code sections which also would skew the results and are therefore omitted.

After preprocessing the raw data, relevant data is filtered and computed. Questions and answers in the data are mixed together and have to be separated and answers have to be linked to their questions. Also, questions in these datasets do not have the *new contributor* indicator attached to them and neither do users. So, the first contribution date and time of users have to be calculated via the creation dates of the questions and answers the user has posted. Then, questions are filtered per user and by whether they are created within the 7-day window after the first contribution of the user. These questions were created during the period where the *new contributor* indicator would have been displayed, in case the questions had been posted before the change, or had been displayed after the change. From these questions, all answers which arrived within the 7-day window are considered for the analysis. Answers which arrived at a later point are excluded as the answerer most likely has not seen the disclaimer shown in figure 3.1. Included answers are then analyzed with Vader and the resulting sentiments are stored. Furthermore, votes to questions of new contributors are counted if they arrived within the 7-day window and count 1 if it is an upvote and -1 if it is a downvote. Moreover, the number of questions new contributors ask, are counted and divided into two classes: 1st-question of a user and follow-up questions of a new contributor.

[4]https://archive.org/download/stackexchange

3.2 Analysis

An interrupted time series (ITS) analysis captures trends before and after a change in a system and fits very well with the question this thesis investigates. ITS can be applied to a large variety of data if the data contains the same kind of data points before and after the change and when the change date and time are known. Bernal, Cummins and Gasparrini published a paper on how ITS works [57]. ITS performs well on medical data, for instance, when a new treatment is introduced ITS can visualize if the treatment improves a condition. For ITS no control group is required and often control groups are not feasible. ITS only works with the before and after data and a point in time where a change was introduced. ITS relies on linear regression and tries to fit a three-segment linear function to the data. The authors also described cases where more than three segments are used but these models quickly raise the complexity of the analysis and for this thesis a three-segment linear regression is sufficient. The three segments are lines to fit the data before and after the change as well as one line to connect the other two lines at the change date. Figure 3.2 shows an example of an ITS. Each segment is captured by a tensor of the following formula $Y_t = \beta_0 + \beta_1 T + \beta_2 X_t + \beta_3 T X_t$, where T represents time as a number, for instance, number of months since the start of data recording, X_t represents 0 or 1 depending on whether the change is in effect, β_0 represents the value at $T = 0$, β_1 represents the slope before the change, β_2 represents the value when the change is introduced, and β_3 represents the slope after the change. Contrary to the basic method explained in [57] where the ITS is performed on aggregated values per month, this thesis performs the ITS on single data points, as the premise that the aggregated values all have the same weight within a certain margin is not fulfilled for sentiment and vote score values. Performing the ITS with aggregated values would skew the linear regression more towards data points with less weight. Single data point fitting prevents this, as weight is taken into account with more data points. To filter out seasonal effects, the average value of all data points with the same month of all years is subtracted from the data points (i.e. subtract the average value of all Januaries from each data point in a January). This thesis uses the least-squares method for regression.

Although the ITS analysis takes data density variability and seasonality into account, there is always a possibility that an unknown factor or event is contained in the data. It is always recommended to do a visual inspection of the data. This thesis

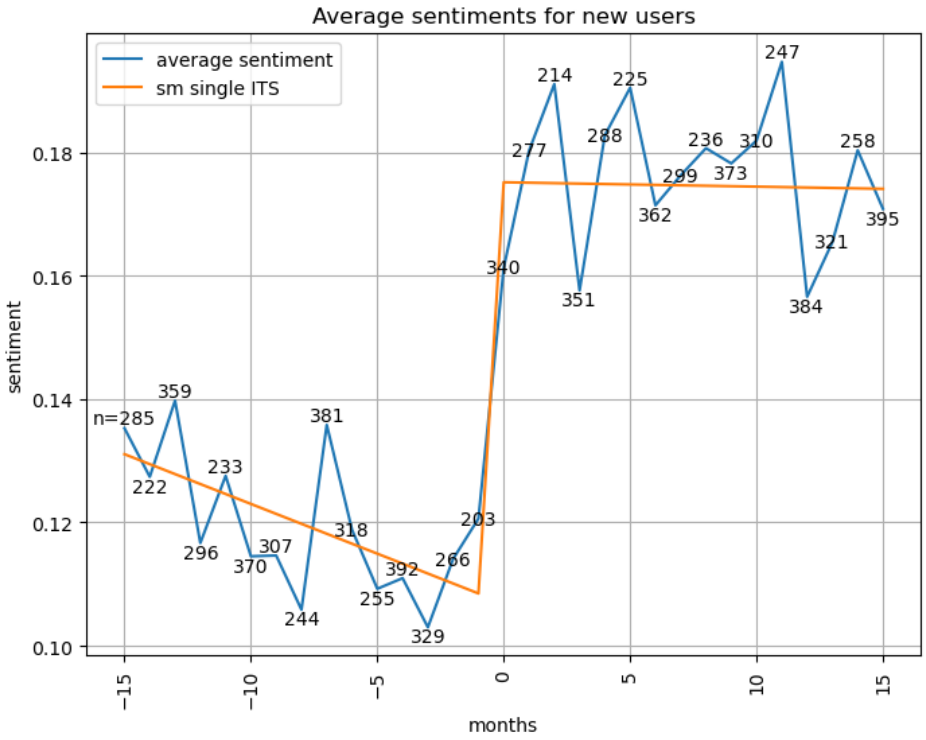


Figure 3.2: An example that visualizes how ITS works. The change of the system occurs at month 0. The blue line shows the average sentiment of fictional answers grouped by month. The numbers attached to the blue line show the number of sentiment values for a given month. The yellow line represents the ITS analysis as a three-segment line. This example shows the expected behavior of the data sets in the following sections.

contains one example where the data density increases so drastically in a particular time span that this form of analysis loses accuracy.

4 Datasets

StackExchange provides complete datasets of its communities for research purposes on archive.org[1]. StackExchange also provides a short guide on how to interpret the provided data, as some data values are strictly numerical and do not convey any meaning without the knowledge of what these values represent. This thesis investigates the largest datasets available and includes the datasets of the following communities:

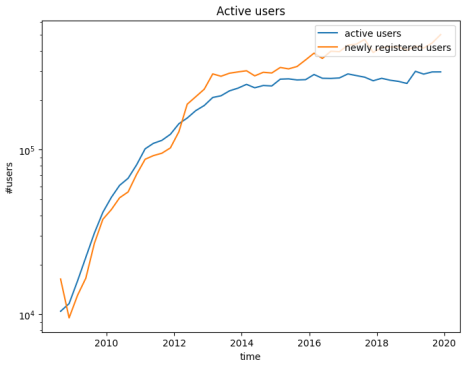
- StackOverflow.com
- math.stackexchange.com
- MathOverflow.net
- AskUbuntu.com
- ServerFault.com
- SuperUser.com
- electronics.stackexchange.com
- stats.stackexchange.com
- tex.stackexchange.com
- unix.stackexchange.com

These datasets are selected due to their size as larger datasets yield more consistent results. Smaller datasets may be too sparse to take any meaningful conclusions. Also, outliers would influence the results more when compared to outliner in bigger datasets. The dataset contain all the necessary data since the creation of the respective community and until the last day of February 2020.

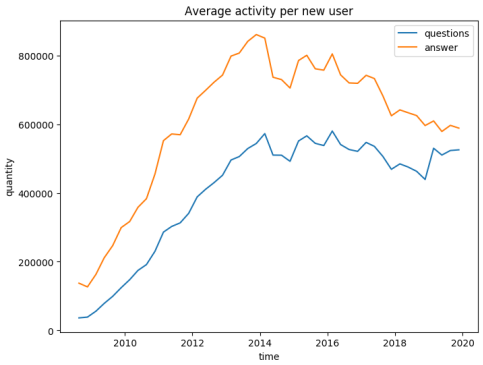
[1] https://archive.org/download/stackexchange

4.1 StackOverflow.com

StackOverflow is a community about software development and programming knowledge and is the largest and oldest community of the StackExchange platform. The community has 11867244 registered users of which 297192 were active between December 2019 and February 2020. Members asked 18699974 questions in total and gave 27981749 answers with an average answer density of 1.496 answers per question. New users asked 2880039 questions with an average of 1.240 questions per new user during their first week after their first contribution.



(a) Active users with activity in the last 3 months

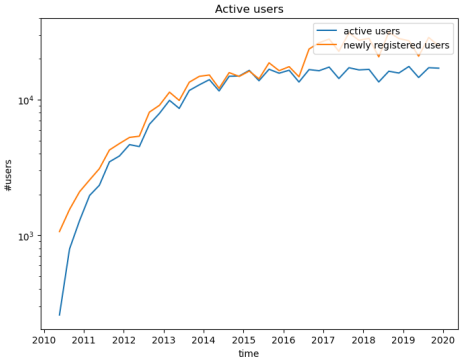


(b) Questions and answers counts over time

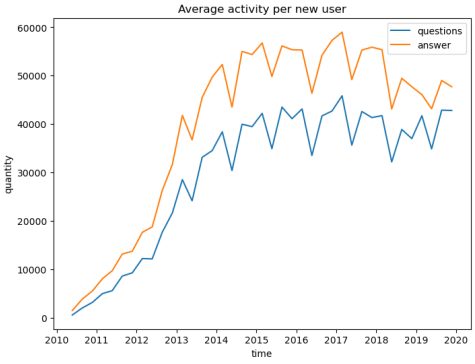
4.2 math.stackexchange.com

"Mathematics Stack Exchange is a question and answer site for people studying math at any level and professionals in related fields."[2] The community has 624671 registered users of which 17074 were active between December 2019 and February 2020. Members asked 1170938 questions in total and gave 1565188 answers with an average answer density of 1.336 answers per question. New users asked 265704 questions with an average of 1.336 questions per new user during their first week after first contribution.

[2] https://math.stackexchange.com/



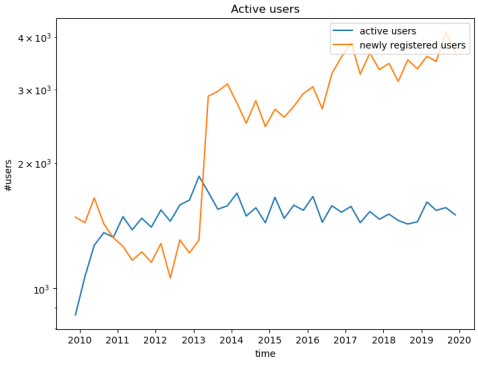
(a) Active users with activity in the last 3 months



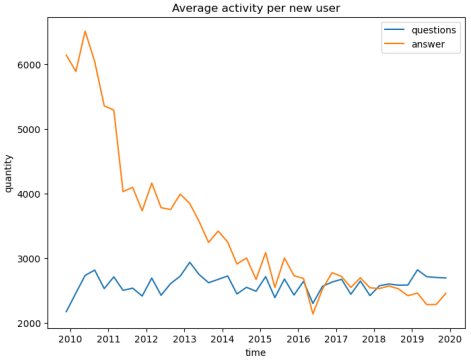
(b) Questions and answers counts over time

4.3 MathOverflow.net

MathOverflow.net is a rather small community for professional mathematicians. The community has 105471 registered users of which 1501 were active between December 2019 and February 2020. Members asked 108083 questions in total and gave 144918 answers with an average answer density of 1.34 answers per question. New users asked 23746 questions with an average of 1.131 questions per new user during their first week after first contribution.



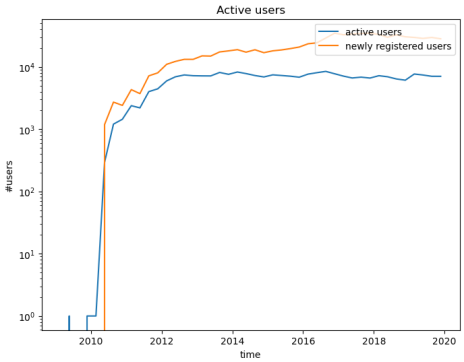
(a) Active users with activity in the last 3 months



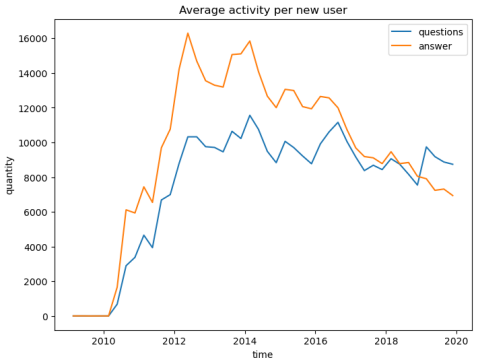
(b) Questions and answers counts over time

4.4 AskUbuntu.com

AskUbuntu.com is a rather small community for Ubuntu users and developers. The community has 783614 registered users of which 7033 were active between December 2019 and February 2020. Members asked 334194 questions in total and gave 418051 answers with an average answer density of 1.25 answers per question. New users asked 157018 questions with an average of 1.101 questions per new user during their first week after first contribution.



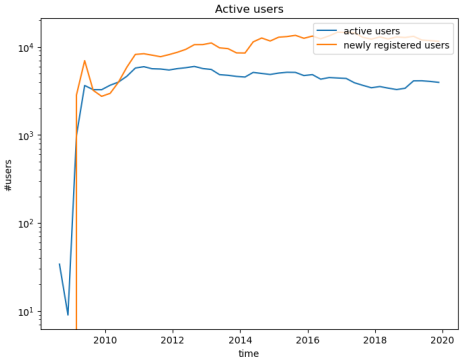
(a) Active users with activity in the last 3 months



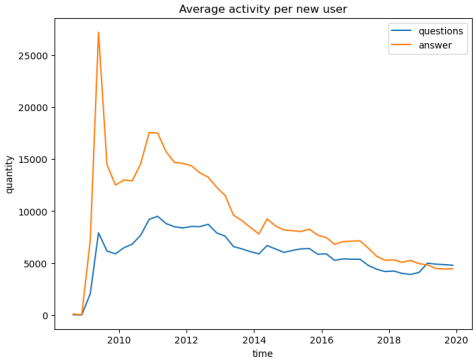
(b) Questions and answers counts over time

4.5 ServerFault.com

ServerFault.com is a rather small community for system and network administrators. The community has 451180 registered users of which 3947 were active between December 2019 and February 2020. Members asked 274564 questions in total and gave 432334 answers with an average answer density of 1.574 answers per question. New users asked 88547 questions with an average of 1.106 questions per new user during their first week after first contribution.



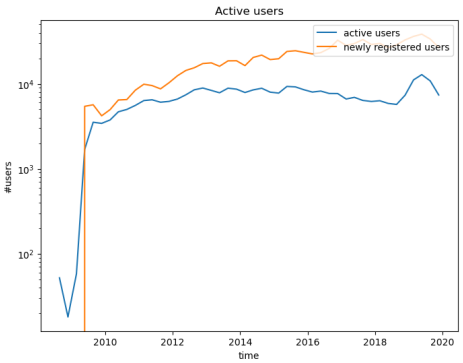
(a) Active users with activity in the last 3 months



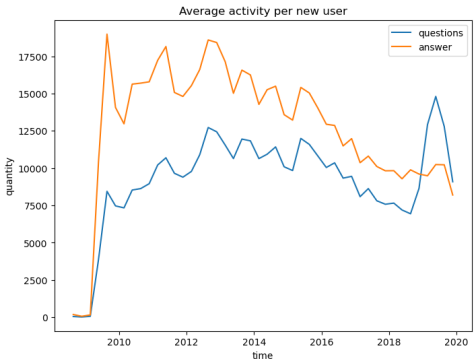
(b) Questions and answers counts over time

4.6 SuperUser.com

SuperUser.com is a rather small community for computer enthusiasts and power users. The community has 861533 registered users of which 7392 were active between December 2019 and February 2020. Members asked 424718 questions in total and gave 587559 answers with an average answer density of 1.383 answers per question. New users asked 161397 questions with an average of 1.085 questions per new user during their first week after first contribution.



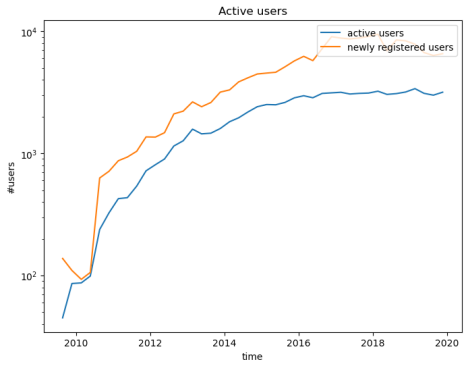
(a) Active users with activity in the last 3 months



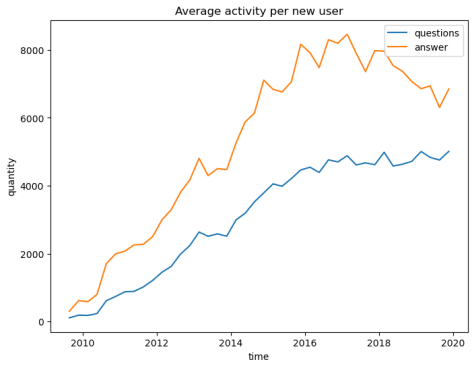
(b) Questions and answers counts over time

4.7 electronics.stackexchange.com

electronics.stackexchange.com is a rather small community for electrical engineering. The community has 184795 registered users of which 3172 were active between December 2019 and February 2020. Members asked 130025 questions in total and gave 221811 answers with an average answer density of 1.705 answers per question. New users asked 47035 questions with an average of 1.126 questions per new user during their first week after first contribution.



(a) Active users with activity in the last 3 months

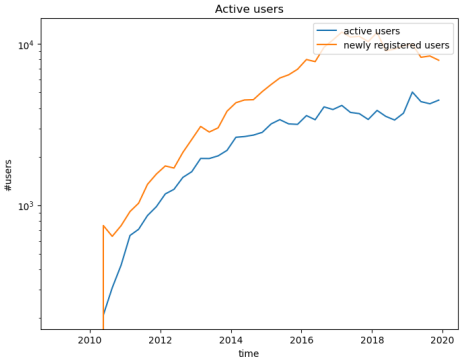


(b) Questions and answers counts over time

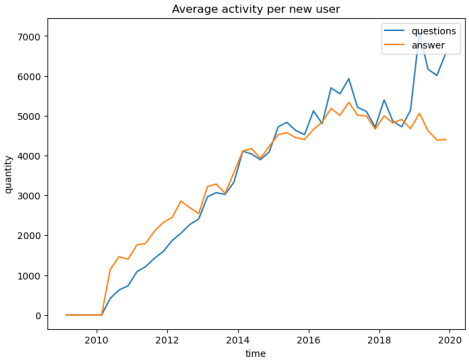
4.8 stats.stackexchange.com (Cross Validated)

"Cross Validated is a question and answer site for people interested in statistics, machine learning, data analysis, data mining, and data visualization."[3] The community has 227032 registered users of which 4485 were active between December 2019 and February 2020. Members asked 151777 questions in total and gave 148046 answers with an average answer density of 0.975 answers per question. New users asked 57636 questions with an average of 1.112 questions per new user during their first week after first contribution.

[3] https://stats.stackexchange.com/



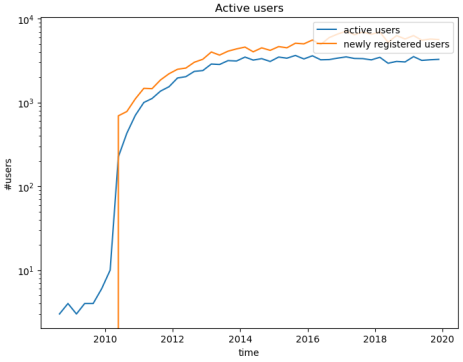
(a) Active users with activity in the last 3 months



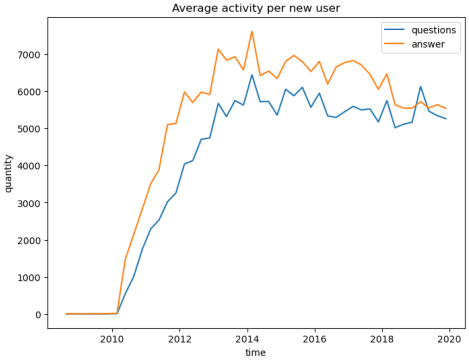
(b) Questions and answers counts over time

4.9 tex.stackexchange.com

tex.stackexchange.com is a rather small community for TEX and related typesetting systems. The community has 171867 registered users of which 3280 were active between December 2019 and February 2020. Members asked 188860 questions in total and gave 227875 answers with an average answer density of 1.206 answers per question. New users asked 59692 questions with an average of 1.191 questions per new user during their first week after first contribution.



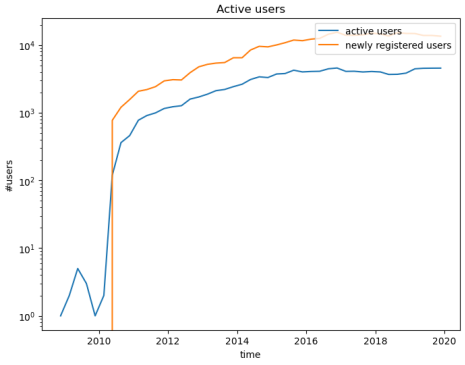
(a) Active users with activity in the last 3 months



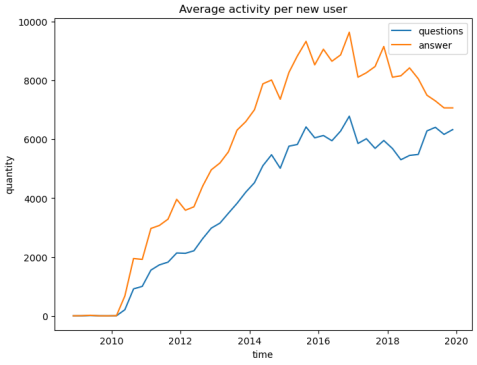
(b) Questions and answers counts over time

4.10 unix.stackexchange.com

unix.stackexchange.com is a rather small community for Linux and Unix-like operating systems. The community has 356498 registered users of which 4565 were active between December 2019 and February 2020. Members asked 174625 questions in total and gave 256007 answers with an average answer density of 1.466 answers per question. New users asked 62437 questions with an average of 1.124 questions per new user during their first week after first contribution.



(a) Active users with activity in the last 3 months



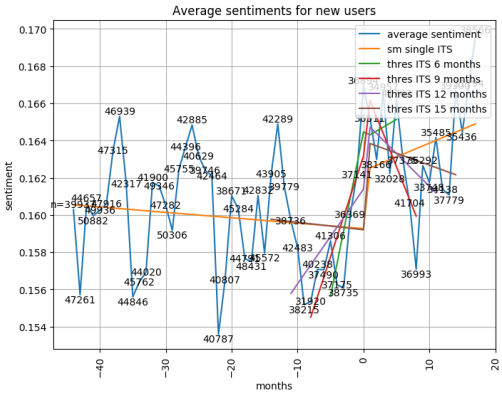
(b) Questions and answers counts over time

5 Results

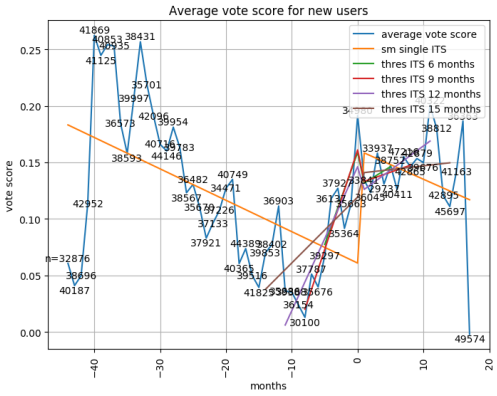
This section shows the results of the experiments described in section 3 on the data sets described in section 4. In the following diagrams, the blue line states the (a) average sentiment of the answers to questions from new contributors, (b) average vote score of questions from new contributors, and (c) the number of 1st and follow-up questions of new contributors. These lines also have numbers attached to it at every data point and each shows (a) the number of answers that formed the sentiment average, and (b) the number of questions that formed the average vote score. The orange line shows ITS analysis as a 3-segment line.

5.1 StackOverflow.com

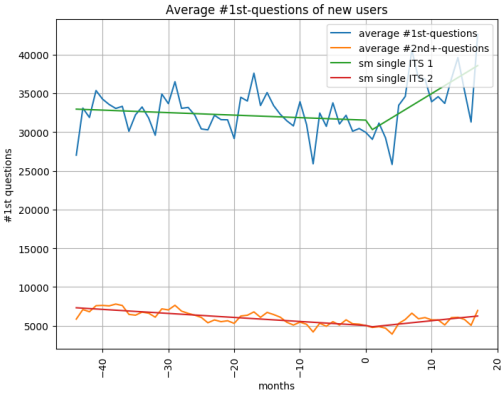
StackOverflow shows a very slight decrease in the average sentiment of time before the change is introduced. When the change occurs the average sentiment jumps up. After the change, the sentiments reach higher levels and keep rising. The average vote score rises right before and stays fairly constant after the change. This indicates that the vote score is not affected by the change. However, the number of questions from new contributors increases after the change while before the change is fairly constant. The number of follow-up questions from new contributors declines before the change and rise after the change.



(a) An interrupted time series analysis of the sentiments of answer to questions created by new contributors on StackOverflow.com



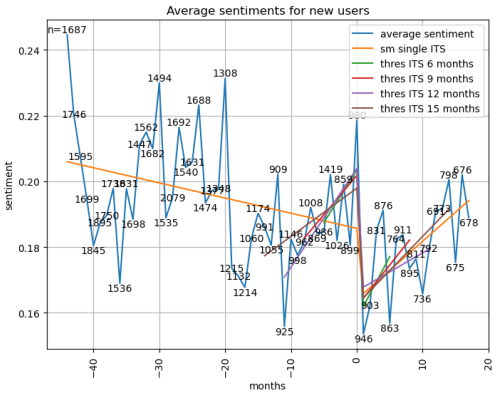
(b) An interrupted time series analysis of the vote score of questions created by new contributors on StackOverflow.com



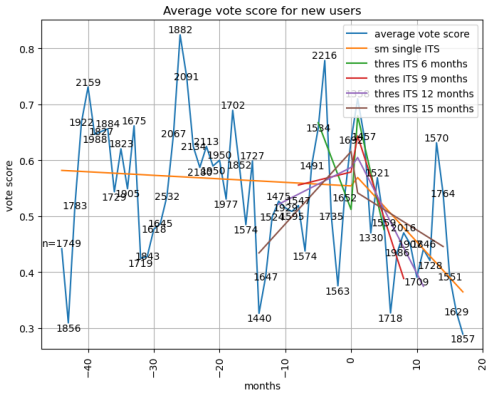
(c) An interrupted time series analysis of the number of questions created by new contributors on StackOverflow.com

5.2 AskUbuntu.com

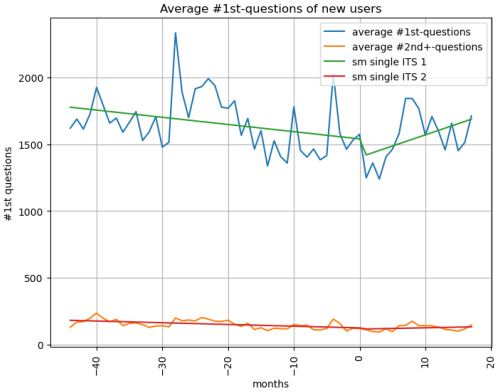
AskUbuntu sees a decrease in average sentiments prior to the change. After the introduction of the change, the regression dips but sentiments keep rising drastically since then. The vote score has a huge range of values prior to and after the change, however, the graph indicates the vote score declines after the change. The number of 1st questions slightly decreases prior to the change and starts rising after the change.



(a) An interrupted time series analysis of the sentiments of answer to questions created by new contributors on AskUbuntu.com



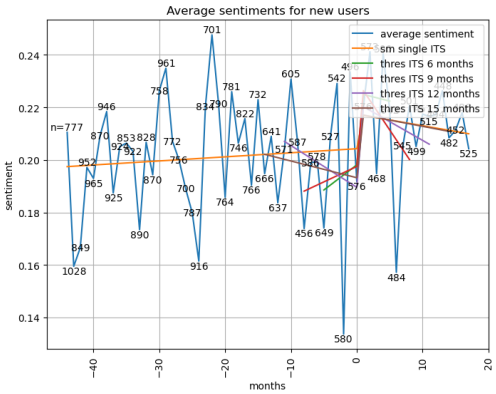
(b) An interrupted time series analysis of the vote score of questions created by new contributors on AskUbuntu.com



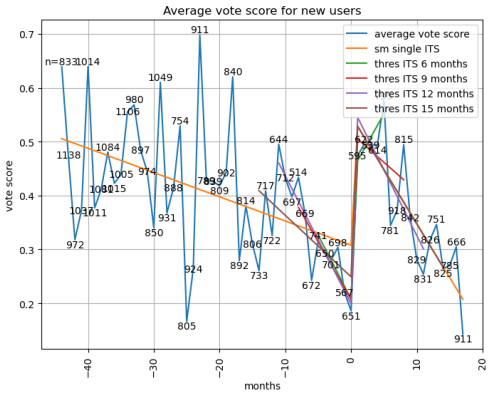
(c) An interrupted time series analysis of the number of questions created by new contributors on AskUbuntu.com

5.3 ServerFault.com

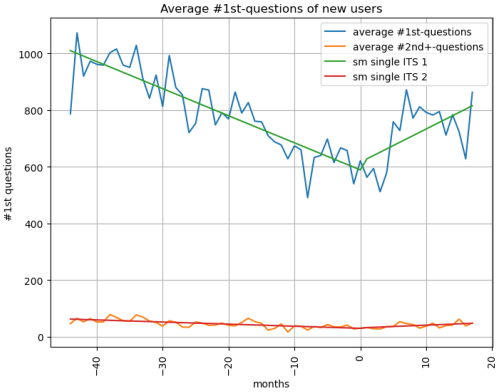
ServerFault shows gradually rising average sentiments prior to the change. At the time of the change, the regression makes a jump upward and the average sentiment decreases slowly afterward. The vote score falls prior to the change, made a huge jump upward, and quickly returns to the levels just prior to the change. The number of 1st questions, however, sees a drastic change. Prior to the change, the number of 1st questions decreases steadily, while after the change the numbers increase at the same pace as they fall prior to the change. The number of follow-up questions also sees the same course direction, falling prior and raising after the change.



(a) An interrupted time series analysis of the sentiments of answer to questions created by new contributors on ServerFault.com



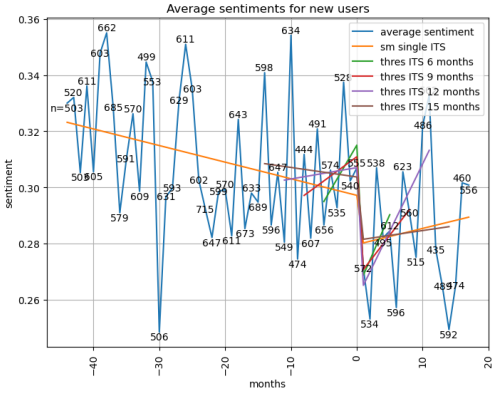
(b) An interrupted time series analysis of the vote score of questions created by new contributors on ServerFault.com



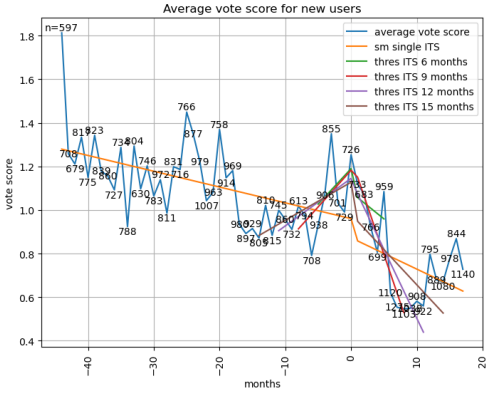
(c) An interrupted time series analysis of the number of questions created by new contributors on ServerFault.com

5.4 stats.stackexchange.com

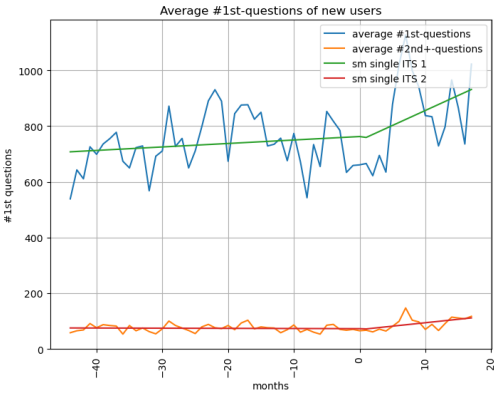
On stats.stackexchange.com the average sentiment decreases steadily prior to the change. The regression dips when the change is introduced. However, the average sentiment after the change indicates a slight upward trend. The vote score also decreases prior to the change but does not recover afterward. However, the number of 1st questions and follow-up questions rise prior to the change and increase even faster after the change.



(a) An interrupted time series analysis of the sentiments of answer to questions created by new contributors on stats.stackexchange.com



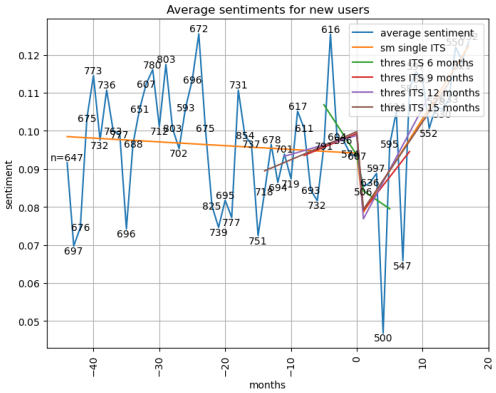
(b) An interrupted time series analysis of the vote score of questions created by new contributors on stats.stackexchange.com



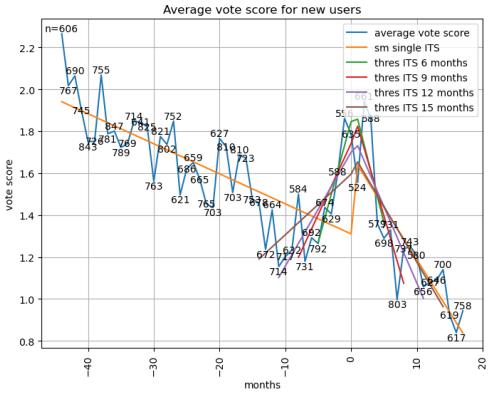
(c) An interrupted time series analysis of the number of questions created by new contributors on stats.stackexchange.com

5.5 tex.stackexchange.com

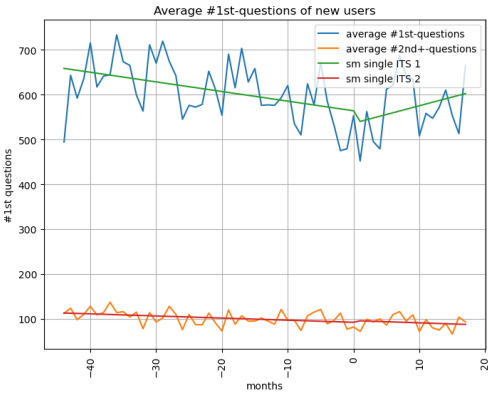
On tex.stackexchange.com the average sentiment is low compared to the other investigated data sets. Prior to the change the average sentiment only slightly decreases. When the change is introduced the regression takes a dip down and after the change, the average sentiment increases drastically. Future data will be required to see if this upward trend continues or evens out. In stark contrast, the vote score shows a downward trend, although there is a short window around the change date where vote scores are higher compared to before and after the change. The number of 1st questions has a downward trend before the change and an upward trend afterward. The downward trend of the number of follow-up questions is uninterrupted by the change.



(a) An interrupted time series analysis of the sentiments of answer to questions created by new contributors on tex.stackexchange.com



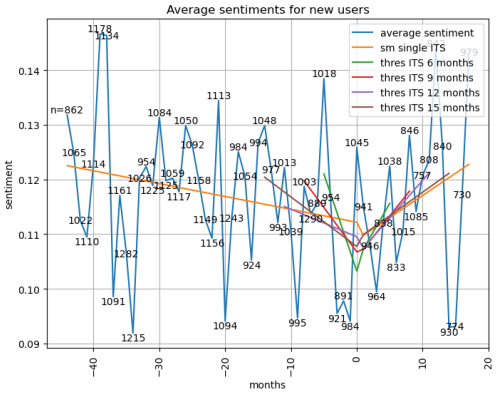
(b) An interrupted time series analysis of the vote score of questions created by new contributors on tex.stackexchange.com



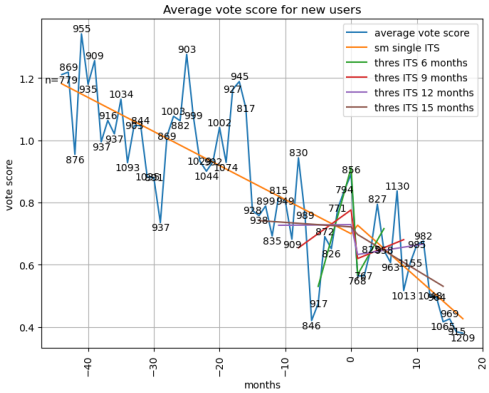
(c) An interrupted time series analysis of the number of questions created by new contributors on tex.stackexchange.com

5.6 unix.stackexchange.com

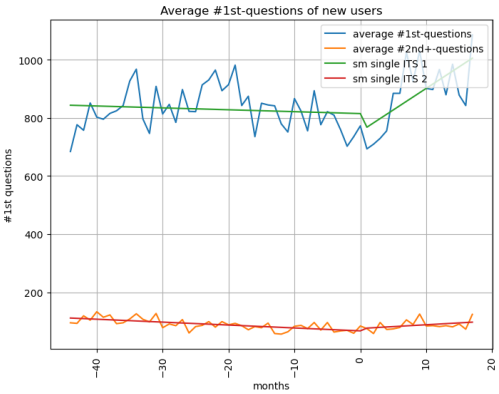
On unix.stackexchange.com the average sentiment decreases prior to the change. When the change is introduced the regression takes a small dip down, however, the average sentiment increases fast after the change. The vote score shows a continuous downward trend and the number of 1st and follow-up questions fall slightly prior to the change and increase afterward.



(a) An interrupted time series analysis of the sentiments of answer to questions created by new contributors on unix.stackexchange.com



(b) An interrupted time series analysis of the vote score of questions created by new contributors on unix.stackexchange.com

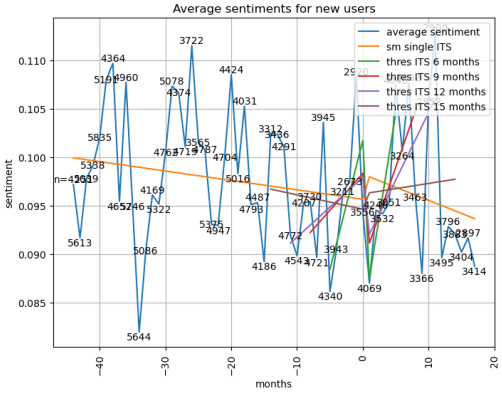


(c) An interrupted time series analysis of the number of questions created by new contributors on unix.stackexchange.com

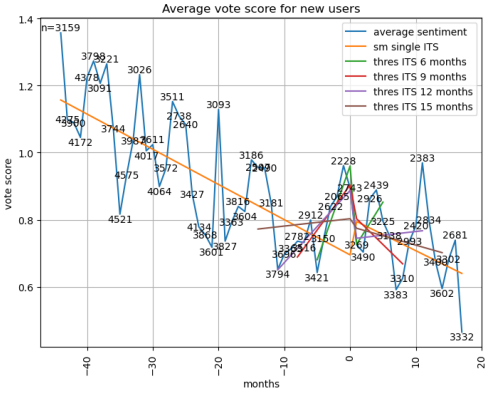
More than half of the commiunities show befits from the change. The number of first questions increase in all of the 6 previously shown communities. Also, for most of these communities the number of follow-up questions increased too. Furthermore, the sentiment ITS shows an improvement in all except 1 community. The vote score analysis yielded no meaningful results for these communities. The vote score does not change with the introduction of Stackexchange' policy, with the exception of ServerFault, however, the increase in the vote score did not last for long.

5.7 math.stackexchange.com

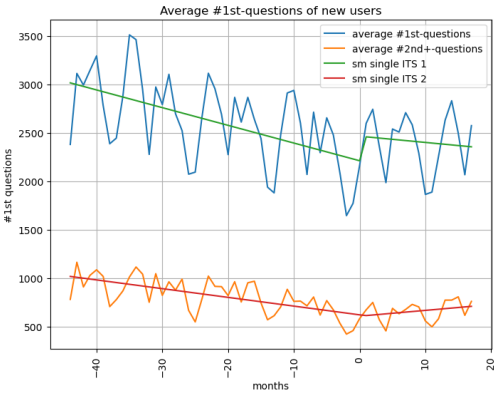
The math.stackexchange.com community shows a decrease in average sentiments, vote score, and the number of questions prior to the change. The measurements make a small jump upward when the change is introduced, however, they continue their downward trend after the introduction of the change. Only the number of follow-up questions stabilizes and begins to increase after the change.



(a) An interrupted time series analysis of the sentiments of answer to questions created by new contributors on math.stackexchange.com



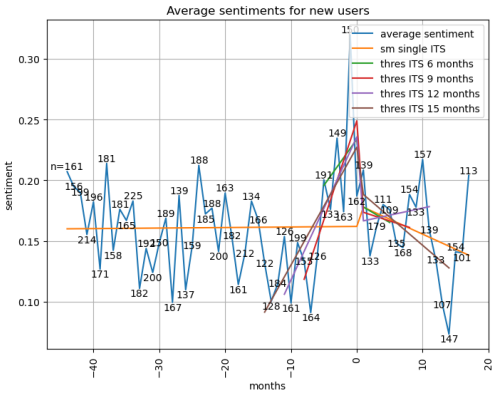
(b) An interrupted time series analysis of the vote score of questions created by new contributors on math.stackexchange.com



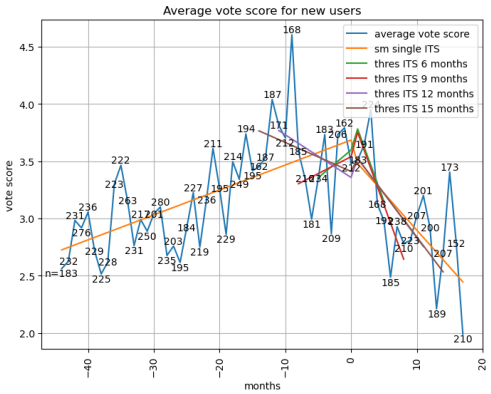
(c) An interrupted time series analysis of the number of questions created by new contributors on math.stackexchange.com

5.8 MathOverflow.net

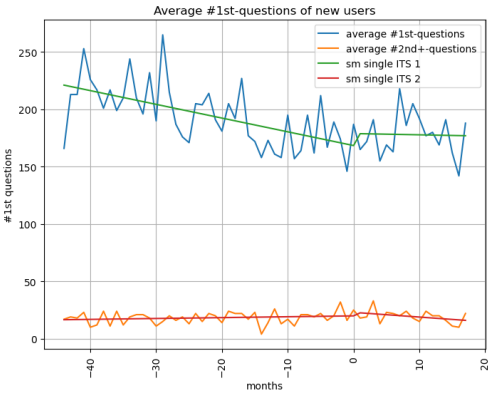
MathOverflow shows a constant regression before the change, however, average sentiments are low at about 10 months before the change and spike high directly before the change. When the change is introduced the regression makes a small jump up and decreases thereafter. The votes score steadily increases prior to the change and then quickly returns to the level from 3 years before the change. The number of 1st questions falls prior to the change and stabilizes thereafter. This data set is sparse compared to the other datasets. Also, the vote scores are high compared to other datasets.



(a) An interrupted time series analysis of the sentiments of answer to questions created by new contributors on MathOverflow.net



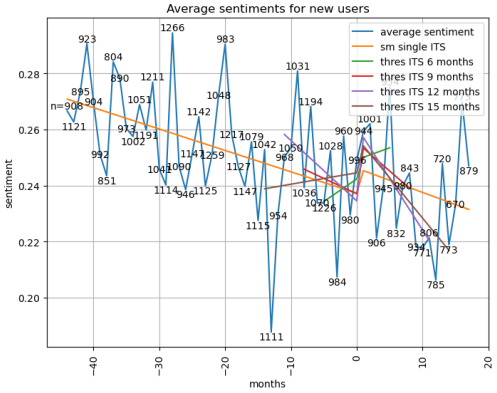
(b) An interrupted time series analysis of the vote score of questions created by new contributors on MathOverflow.net



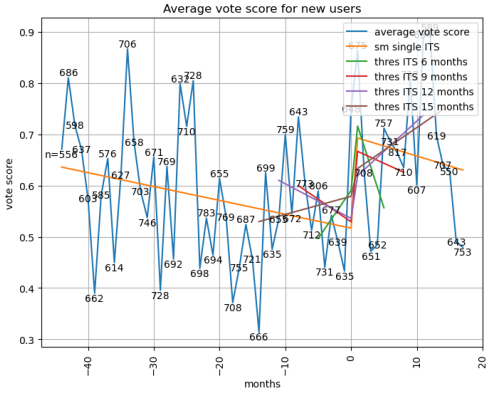
(c) An interrupted time series analysis of the number of questions created by new contributors on MathOverflow.net

5.9 electronics.stackexchange.com

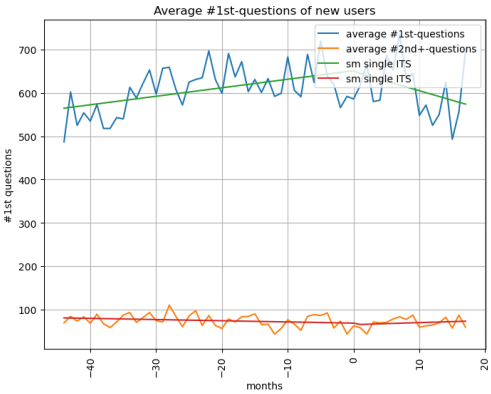
On electronics.stackexchange.com the average sentiment and votes decrease continuously prior to the change. At the change date, the regression makes a little jump upward but the trend from before the change continues afterward. Similarly to SuperUser, the average sentiment recovers at about 12 months after the change is introduced and future data will be necessary to determine if the recovery is persistent. The number of 1st questions rises continuously prior to the change and decreases thereafter. The number of follow-up questions falls slightly prior to the change and stabilizes afterward.



(a) An interrupted time series analysis of the sentiments of answer to questions created by new contributors on electronics.stackexchange.com



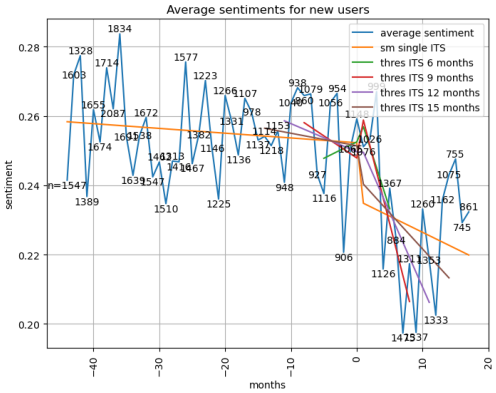
(b) An interrupted time series analysis of the vote score of questions created by new contributors on electronics.stackexchange.com

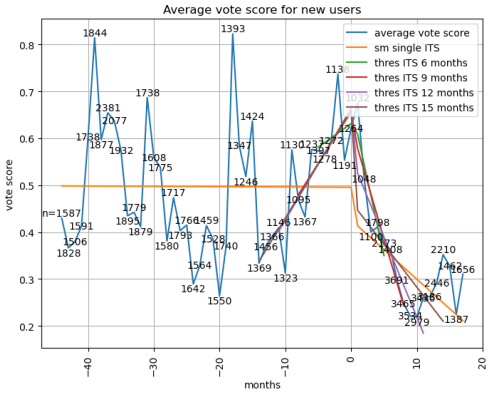


(c) An interrupted time series analysis of the number of questions created by new contributors on electronics.stackexchange.com

5.10 SuperUser.com

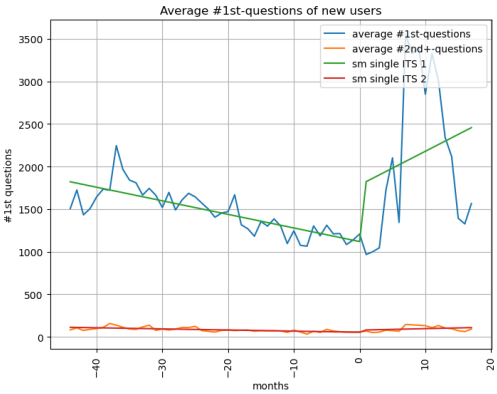
SuperUser shows only sightly decreasing average sentiment and vote score up to the change. At the change time the regressions take a dip down and the regression shows a downward trend after the change. Indeed the average sentiments and vote score dipped considerably when the change is introduced. The average sentiment recovers about 13 months later, while the vote score does not recover as well. The number of 1st questions decreases prior to the change and then goes through the roof indicating a huge wave of new users. This drastic influx of new users may explain the crash of the average sentiment and vote score that occurs at the same time. Data available in the future will show if the recovery is persistent.





(a) An interrupted time series analysis of the sentiments of answer to questions created by new contributors on SuperUser.com

(b) An interrupted time series analysis of the vote score of questions created by new contributors on SuperUser.com



(c) An interrupted time series analysis of the number of questions created by new contributors on SuperUser.com

The 4 previously mentioned communities do not profit from the change. Although some communities improve in one statistic, they do not improve accross the field as shown in the other 6 communities. The 1st question statistic decreases in all 4 communities. With the exception of math.stackexchange.com, all of these communities do not improve in the followup question statistic. In all communities the vote score is on a (worse) downward trend after the change. Also, the sentiment values are decreasing after the change.

When looking at the results of SuperUser, the community stands out and shows interesting results. After about 6 mouths after the change the community the number of 1st questions triple. This level of new questions continues for 7 months

before the the number go down towards the previous levels. In the same time frame the vote score and sentiment take a significant dive. After that the sentiment returns almost to the previous level while the vote score only increases mildly. However, this sudden increase in 1st questions and therefore users is not related to the change this thesis investivates.

Summarizing, the change introduced by StackExchange clearly improved the engagement in 6 of the 10 investigated communities. Sentiment, vote score, and number (1st and follow-up questions) rose as a result. The other 4 communities do not profit from the change. Although, many statistics jump up to a higher level the downward trends are not stopped. The statistics of SuperUser show a large influx of new users about 6 months after the change sending the sentiment and vote score on a deep dive and with the decrease in new users they raise again. However, this event is not related to the change but the magnitude of the huge change in new user numbers renders the analysis uncomparable.

6 Discussion

The ITS analysis of the investigated communities shows mixed results. Some communities show an increase in sentiment while others are not affected at all or show a decrease in sentiment. The StackOverflow community has a fairly stable average sentiment before the change. The average sentiment jumps to a higher level and keeps rising after the change is introduced. Furthermore, the number of 1st questions from new contributors starts rising drastically after the change while prior levels stagnate. Also, the follow-up questions start increasing slightly. The votes score trend takes a new direction 9 months before the change and is unrelated to it. The change has a positive effect on the StackOverflow community. Beside StackOverflow, 5 other communities seem to profit from the change: AskUbuntu, ServerFault, stats.stackexchange.com, tex.stackexchange.com, and unix.stackexchange.com. AskUbuntu shows an interesting zig-zag pattern in the average sentiment graph. Also, the average sentiment falls before the change and raises thereafter, indicating that the change works for this community. However, further data is needed to see if the zig-zag pattern repeats itself. The number of 1st questions starts increasing again after the change stopping the downward trend before that. On stats.stackexchange.com the average sentiment falls before the change but since the change, the downward trend stops and the sentiment starts to rise slowly, suggesting the change has a positive effect on the community. This is supported by the increase in the number of 1st and followup questions by new contributors. The vote score takes a dip after the change but starts to recover after 12 months which could be the result of another change. In the tex.stackexchange.com community sentiments are stable before the change and show a stark rising pattern after the change. The change seems to work for this community but future data will be necessary to see if the rising pattern continues in the shown manner. The votes score ITS does not fit the model and values before and after the change indicate a linear downward trend. However, the number of 1st questions increases slightly after the change while the prior trend shows a decreasing development. unix.stackexchange.com also shows a decreasing pattern prior and a rising pattern

after the change. The vote score analysis shows a fairly linear downward trend before and after the change and is not affected by it. However, the number of 1st questions by new contributors starts to drastically increase while before the change the levels are constant, indicating this community also profits from the change. On ServerFault the sentiment rises gradually before the change, jumps upward by a small value when the change is introduced and the sentiment falls slowly thereafter but the levels are pretty stable over the analyzed period. The vote scores show the change has a huge impact on the community. The previously decreasing trend jumps up by a large amount. However, the vote score rapidly returns to levels right before the change. Contrary, the number of first questions turns direction and starts increasing at the same rate it is falling previously.

The other communities do not seem to profit from the change so clearly. The average sentiment stays constant on MathOverflow before the change and decreases afterward. However, the sentiment levels start increasing six months before the change and are unrelated, indicating the sentiment values are not particularly affected by the change. The vote score is steadily increasing before the change and the crashes down shortly after the change. However, the vote score is very high compared to other communities. The number of 1st questions stabilizes after the change compared to the slight downward previously. math.stackexchange.com shows a downward trend before and after the change for sentiment and vote score. The sentiment ITS is particularly affected by the low sentiment values at the end and future data is required to determine if this trend continues. However, the number of 1st questions stabilizes a bit after changes and follow up questions even see and a slight increase after the change. The electronics.stackexchange.com community has a similar pattern for the sentiment value and vote scores compared to math.stackexchange.com. However, the sentiment values seem to recover after about 12 months and future data is required to see if the rise at the end of the period is a long term trend. The rising number of first questions of new contributors stops at the change date and transition into a decreasing pattern. SuperUser shows an odd pattern. The average sentiment values and votes scores are stable before the change and decrease dramatically shortly afterward. However, the sentiment recovers after 12 months. The ITS model chosen in this thesis is not able to capture the apparent pattern. However, the number of 1st question skyrockets indicating a huge influx of new users. The time frames of the falling sentiment values and vote scores, and the rising number of first questions overlap, indicating the huge influx of new users is responsible for the falling patterns.

By and large, the change introduced by the StackExchange team has a clear positive effect on more than half of the investigated communities. Two of the communities, SuperUser and stats.stackexchange.com, have a delayed temporary decrease in sentiment which recovers after about 12 months, which may be attributable to the larger influx of new contributors. The selected ITS model is not designed to capture the sentiment pattern of these communities. math.stackexchange.com is not really affected by the change, although the number of 1st questions stabilized a bit and follow-up questions from new contributors increase again. MathOverflow shows a similar picture.

Some investigated data sets show interesting patterns. StackOverflow shows the clearest results of all the investigated communities and closely resembles the example ITS shown in section 3. The result matches the expectation, that advising answerers to remember the code of conduct when answering questions from new contributors will increase the welcomingness and friendliness of the community, and shows that the change introduced by the StackExchange team works well for this community. The AskUbuntu community shows an interesting zig-zag pattern where sentiment gradually rises over time and then falls abruptly.

The average sentiment of the StackOverflow community is the most stable in terms of deviation from the regression. This is expected as StackOverflow is the largest community by far and has the most questions created by newcomers. On the other hand, MathOverflow is the sparsest community and has the least amount of questions from new contributors. The level of the average sentiment also varies greatly between communities. stats.stackexchange.com has the highest level of average sentiment compared to the other communities, whereas, tex.stackexchange.com has the lowest level average sentiment. MathOverflow has the highest level of vote scores by far. Also, in most communities, the number of questions from new contributors slowly decreases over time. This may be a result of the filling of gaps in the knowledge repository over time.

7 Conclusion

The change introduced by the StackExchange team produced desired results in more than half of the investigated communities. The results of the StackOverflow community most closely resembles the expectation of improving the welcomingness. AskUbuntu, ServerFault, stats.stackexchange.com, tex.stackexchange.com, and unix.stackexchange also profit from this change. MathOverflow, SuperUser, math.stackexchange.com, and electronics.stackexchange.com do not profit as much from the change and show not an increase but decrease or continuation in the decrease of sentiment. However, the falling number of questions from new contributors stabilized a bit for the math communities and the vote score increased for electronics.stackexchange.com. SuperUser saw a huge influx of new contributors shortly after the change who asked a lot of questions and dropping the sentiment and vote score value during that period.

Bibliography

[1] Clayton J Hutto and Eric Gilbert. 'Vader: A parsimonious rule-based model for sentiment analysis of social media text'. In: *Eighth international AAAI conference on weblogs and social media*. 2014.

[2] Dana Movshovitz-Attias et al. 'Analysis of the reputation system and user contributions on a question answering website: Stackoverflow'. In: *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. IEEE. 2013, pp. 886–893.

[3] Lena Mamykina et al. 'Design lessons from the fastest q&a site in the west'. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2011, pp. 2857–2866.

[4] Rogier Slag, Mike de Waard and Alberto Bacchelli. 'One-day flies on stackoverflow-why the vast majority of stackoverflow users only posts once'. In: *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*. IEEE. 2015, pp. 458–461.

[5] Denae Ford et al. 'We Don't Do That Here: How Collaborative Editing with Mentors Improves Engagement in Social Q&A Communities'. In: *CHI*. 2018.

[6] Blerina Bazelli, Abram Hindle and Eleni Stroulia. 'On the personality traits of stackoverflow users'. In: *2013 IEEE international conference on software maintenance*. IEEE. 2013, pp. 460–463.

[7] Amiangshu Bosu et al. 'Building reputation in stackoverflow: an empirical investigation'. In: *2013 10th Working Conference on Mining Software Repositories (MSR)*. IEEE. 2013, pp. 89–92.

[8] Stav Yanovsky et al. 'One Size Does Not Fit All: Badge Behavior in Q&A Sites'. In: *UMAP*. 2019.

[9] Tomasz Kusmierczyk and Manuel Gomez-Rodriguez. 'On the Causal Effect of Badges'. In: *WWW*. 2018.

[10] Ashton Anderson et al. 'Steering user behavior with badges'. In: *WWW*. 2013.

[11] Nicole Immorlica, Greg Stoddard and Vasilis Syrgkanis. 'Social status and badge design'. In: *Proceedings of the 24th international conference on World Wide Web*. 2015, pp. 473–483.

[12] Yla R Tausczik and James W Pennebaker. 'Predicting the perceived quality of online mathematics contributions from users' reputations'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2011, pp. 1885–1888.

[13] Gang Wang et al. 'Wisdom in the social crowd: an analysis of quora'. In: *Proceedings of the 22nd international conference on World Wide Web*. 2013, pp. 1341–1352.

[14] Zhiyuan Lin et al. 'Better when it was smaller? Community content and behavior after massive growth'. In: *Eleventh International AAAI Conference on Web and Social Media*. 2017.

[15] Eshwar Chandrasekharan et al. 'You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech'. In: *Proceedings of the ACM on Human-Computer Interaction* 1.CSCW (2017), pp. 1–22.

[16] Jiang Bian et al. 'Finding the right facts in the crowd: factoid question answering over social media'. In: *Proceedings of the 17th international conference on World Wide Web*. 2008, pp. 467–476.

[17] Imrul Kayes et al. 'The social world of content abusers in community question answering'. In: *Proceedings of the 24th international conference on world wide web*. 2015, pp. 570–580.

[18] Ramtin Yazdanian et al. 'Eliciting New Wikipedia Users' Interests via Automatically Mined Questionnaires: For a Warm Welcome, Not a Cold Start'. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 13. 01. 2019, pp. 537–547.

[19] Christoph Treude, Ohad Barzilay and Margaret-Anne Storey. 'How do programmers ask and answer questions on the web?(NIER track)'. In: *Proceedings of the 33rd international conference on software engineering*. 2011, pp. 804–807.

[20] Pierre N Robillard. 'The role of knowledge in software development'. In: *Communications of the ACM* 42.1 (1999), pp. 87–92.

[21] Robert E Kraut and Paul Resnick. *Building successful online communities: Evidence-based social design*. Mit Press, 2012.

[22] Panagiotis G Ipeirotis and Evgeniy Gabrilovich. 'Quizz: targeted crowdsourcing with a billion (potential) users'. In: *Proceedings of the 23rd international conference on World wide web*. 2014, pp. 143–154.

[23] Igor Steinmacher et al. 'Social barriers faced by newcomers placing their first contribution in open source software projects'. In: *Proceedings of the 18th ACM conference on Computer supported cooperative work & social computing*. 2015, pp. 1379–1392.

[24] David Allen. 'Do organizational socialization tactics influence newcomer embeddedness and turnover?' In: *Journal of Management* (2006).

[25] Blair Nonnecke, Dorine Andrews and Jenny Preece. 'Non-public and public online community participation: Needs, attitudes and behavior'. In: *Electronic Commerce Research* 6.1 (2006), pp. 7–20.

[26] Peter Kollock and Marc Smith. 'Managing the virtual commons'. In: *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives* (1996), pp. 109–128.

[27] Merrill Morris and Christine Ogan. 'The Internet as mass medium'. In: *Journal of Computer-Mediated Communication* 1.4 (1996), JCMC141.

[28] Denae Ford et al. 'Paradise unplugged: Identifying barriers for female participation on stack overflow'. In: *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*. 2016, pp. 846–857.

[29] Bogdan Vasilescu, Andrea Capiluppi and Alexander Serebrenik. 'Gender, representation and online participation: A quantitative study'. In: *Interacting with Computers* 26.5 (2014), pp. 488–511.

[30] Paul A David and Joseph S Shapiro. 'Community-based production of open-source software: What do we know about the developers who participate?' In: *Information Economics and Policy* 20.4 (2008), pp. 364–398.

[31] Jacob Clark Blickenstaff*. 'Women and science careers: leaky pipeline or gender filter?' In: *Gender and education* 17.4 (2005), pp. 369–386.

[32] Catherine Hill, Christianne Corbett and Andresse St Rose. *Why so few? Women in science, technology, engineering, and mathematics*. ERIC, 2010.

[33] Alicia Iriberri and Gondy Leroy. 'A life-cycle perspective on online community success'. In: *ACM Computing Surveys (CSUR)* 41.2 (2009), pp. 1–29.

[34] Diane Maloney-Krichmar and Jenny Preece. 'A multilevel analysis of sociability, usability, and community dynamics in an online health community'. In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 12.2 (2005), pp. 201–232.

[35] Brian Butler et al. 'Community effort in online groups: Who does the work and why'. In: *Leadership at a distance: Research in technologically supported work* 1 (2002), pp. 171–194.

[36] Mark Ginsburg and Suzanne Weisband. 'A framework for virtual community business success: The case of the internet chess club'. In: *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the.* IEEE. 2004, 10–pp.

[37] Edward E Lawler III. *Rewarding excellence: Pay strategies for the new economy.* Jossey-Bass, 2000.

[38] Ivan Srba and Maria Bielikova. 'Why is stack overflow failing? preserving sustainability in community question answering'. In: *IEEE Software* 33.4 (2016), pp. 80–89.

[39] Ashton Anderson et al. 'Discovering value from community activity on focused question answering sites: a case study of stack overflow'. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.* 2012, pp. 850–858.

[40] Huseyin Cavusoglu, Zhuolun Li and Ke-Wei Huang. 'Can gamification motivate voluntary contributions? The case of StackOverflow Q&A community'. In: *Proceedings of the 18th ACM conference companion on computer supported cooperative work & social computing.* 2015, pp. 171–174.

[41] Zhuolun Li, Ke-Wei Huang and Huseyin Cavusoglu. 'Quantifying the impact of badges on user engagement in online Q&A communities'. In: (2012).

[42] Luca Ponzanelli et al. 'Improving low quality stack overflow post detection'. In: *2014 IEEE international conference on software maintenance and evolution.* IEEE. 2014, pp. 541–544.

[43] Justin Cheng, Cristian Danescu-Niculescu-Mizil and Jure Leskovec. 'Antisocial behavior in online discussion communities'. In: *Proceedings of the International AAAI Conference on Web and Social Media.* Vol. 9. 1. 2015.

[44] James W Pennebaker, Martha E Francis and Roger J Booth. 'Linguistic inquiry and word count: LIWC 2001'. In: *Mahway: Lawrence Erlbaum Associates* 71.2001 (2001), p. 2001.

[45] James W Pennebaker et al. 'The Development and Psychometric Properties of LIWC2007'. In: ().

[46] James W Pennebaker et al. *The development and psychometric properties of LIWC2015*. Tech. rep. 2015.

[47] Philip J Stone, Dexter C Dunphy and Marshall S Smith. 'The general inquirer: A computer approach to content analysis.' In: (1966).

[48] Minqing Hu and Bing Liu. 'Mining and summarizing customer reviews'. In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004, pp. 168–177.

[49] Bing Liu, Minqing Hu and Junsheng Cheng. 'Opinion observer: analyzing and comparing opinions on the web'. In: *Proceedings of the 14th international conference on World Wide Web*. 2005, pp. 342–351.

[50] Erik Cambria et al. 'Senticnet: A publicly available semantic resource for opinion mining.' In: *AAAI fall symposium: commonsense knowledge*. Vol. 10. 0. Citeseer. 2010.

[51] Margaret M Bradley and Peter J Lang. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Tech. rep. Technical report C-1, the center for research in psychophysiology …, 1999.

[52] George A Miller. 'WordNet: a lexical database for English'. In: *Communications of the ACM* 38.11 (1995), pp. 39–41.

[53] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[54] Stefano Baccianella, Andrea Esuli and Fabrizio Sebastiani. 'Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.' In: *Lrec*. Vol. 10. 2010. 2010, pp. 2200–2204.

[55] Cem Akkaya, Janyce Wiebe and Rada Mihalcea. 'Subjectivity word sense disambiguation'. In: *Proceedings of the 2009 conference on empirical methods in natural language processing*. 2009, pp. 190–199.

[56] David McDowall, Richard McCleary and Bradley J Bartos. *Interrupted time series analysis*. Oxford University Press, 2019.

[57] James Lopez Bernal, Steven Cummins and Antonio Gasparrini. ‘Interrupted time series regression for the evaluation of public health interventions: a tutorial’. In: *International journal of epidemiology* 46.1 (2017), pp. 348–355.

Appendix