Empirical Software Engineering (2024) 29:156
https://doi.org/10.1007/s10664-024-10544-7

Exploring user privacy awareness on GitHub: an empirical study

Costanza Alfieri[1] · **Juri Di Rocco**[1] · **Paola Inverardi**[2] · **Phuong T. Nguyen**[1]

Accepted: 5 September 2024 / Published online: 27 September 2024

Abstract

GitHub provides developers with a practical way to distribute source code and collaboratively work on common projects. To enhance account security and privacy, GitHub allows its users to manage access permissions, review audit logs, and enable two-factor authentication. However, despite the endless effort, the platform still faces various issues related to the privacy of its users. This paper presents an empirical study delving into the GitHub ecosystem. Our focus is on investigating the utilization of privacy settings on the platform and identifying various types of sensitive information disclosed by users. Leveraging a dataset comprising 6,132 developers, we report and analyze their activities by means of comments on pull requests. Our findings indicate an active engagement by users with the available privacy settings on GitHub. Notably, we observe the disclosure of different forms of private information within pull request comments. This observation has prompted our exploration into sensitivity detection using a large language model and BERT, to pave the way for a personalized privacy assistant. Our work provides insights into the utilization of existing privacy protection tools, such as privacy settings, along with their inherent limitations. Essentially, we aim to advance research in this field by providing both the motivation for creating such privacy protection tools and a proposed methodology for personalizing them.

Keywords Empirical study · Experience report · Sensitivity detection · Privacy · Large language models · BERT · Privacy profile

Communicated by: Fabio Calefato, Hourieh Khalajzadeh and Igor Steinmacher

This article belongs to the Topical Collection: *Special Issue on CHASE 2023*.

✉ Costanza Alfieri
costanza.alfieri@student.univaq.it

Juri Di Rocco
juri.dirocco@univaq.it
https://jdirocco.github.io

Paola Inverardi
paola.inverardi@gssi.it
https://gssi.it/institute/organization/item/220-inverardi-paola

Phuong T. Nguyen
phuong.nguyen@univaq.it
https://www.disim.univaq.it/ThanhPhuong

[1] DISIM - University of L'Aquila, L'Aquila, Italy

[2] Gran Sasso Science Institute, L'Aquila, Italy

1 Introduction

In recent years, privacy in the digital world has become a major concern. Regulations like the EU General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), or the UK's Data Protection Act have been ratified to regulate the (ab-)use of sensitive data (Voigt and Von dem Bussche 2017; Pardau 2018; Jay 2000). While being undeniably valuable, it has become clear that regulations alone may not suffice to ensure robust protection of user privacy. Indeed, software platforms provide mechanisms that allow users to set their privacy preferences and that, together with regulation should offer a wider shield to user privacy.

GitHub[1] is a platform for software developers where a significant amount of professional collaboration and coding takes place. As a matter of fact, it holds vast amounts of data that can be privacy sensitive for its users. GitHub users are first class inhabitants of the digital world and may be considered experts. However, ensuring that they are aware of the sensitivity of personal information that they leave on the platform and that they have control over who can see this data and how it's used, is crucial for maintaining trust in the platform and protecting users' interests and safety. Indeed, GitHub offers to users means to declare which pieces of personal information they are willing to make public setting this way their privacy preferences. However, this is not enough provided that in their daily activity users may leave intentionally or unintentionally breadcrumbs of personal information open to the public.

In this paper, we show that diverse pieces of personal information about a specific user can be uncovered on GitHub, and this information may disclose data that are hidden in the user's privacy settings. In the study, we extend upon concepts introduced in previous research. In particular, we take privacy settings provided on GitHub as a declaration of intent of the user that defines their privacy profile (Inverardi et al. 2023; Migliarini et al. 2020). Our exploration focuses on determining whether these privacy settings are actively utilized by users and how. We leverage this information to compare it with the actual behaviors of users, particularly what they disclose in their comments. Our objective is to enhance our understanding of the effectiveness of privacy settings and explore any potential correlation with the users' observable behaviors. Our research offers insights into the utilization of existing privacy protection tools, such as privacy settings, along with their inherent limitations. Given the interest in developing privacy assistants (e.g., Liu et al. (2016)), our research supports this field by providing both the motivation for creating such privacy protection tools and a proposed methodology for personalizing them.

To achieve this, we conducted an analysis of pull_request comments within a subset of GitHub users, namely the *Active users* (see Section 4.1), intending to identify what type of private information can be found on the platform. The choice of restricting to pull_request is motivated by previous studies in Sajadi et al. (2023); Iyer et al. (2019) (see Section 2). Our examination yielded insights into users' family info, moral values, workplace details, travel habits, time zones, and more. In conjunction with these findings, we present a dataset comprising both sensitive and non-sensitive comments, each manually labeled into distinct sensitive classes.

In the analysis process, we curated a labeled dataset of pull_requests comments made by the *Active users*. The novelty of this labeled dataset consists in having labeled *"digital behaviour"*, that can be linked to certain privacy profiles. The analysis discovered several discrepancies between the user privacy settings and pieces of information disclosed by the user in some comments. These discrepancies may represent instances of the so-called *"privacy*

[1] https://github.com/

paradox," i.e., inconsistencies between users declared preferences and their actual behavior (Barth and de Jong 2017; Kokolakis 2017), where users appear to be very attentive about their privacy, but eventually in their actual behavior miss to defend their personal data. Alternatively, they may simply indicate a lack of attention/importance of the user for the privacy settings regarding her actual behavior. In both cases, the results show that the way personal privacy is dealt with in GitHub needs to be improved.

To address similar problems, many researchers claim the importance of having automated tools to protect users' privacy in the digital world (Liu et al. 2016; Fukuyama et al. 2021; Autili et al. 2019).

Our work represents a step forward in this direction allowing for the realization of an awareness tool that can assist the users in composing their textual comments in a privacy consistent way with the user profile. A proof of concepts of such a tool is presented in Section 5.

To summarize, in this work we answer the following research questions that help understanding the privacy dynamics on GitHub:

- **RQ**$_1$: *Are the privacy settings provided on GitHub used/adopted by the users? In other terms, do we observe different combinations of these settings, or is there a dominant configuration?* We investigate a large set of developers to find out if there exist differences in their privacy preferences. The use of privacy settings and the potential differences between users' selection indicate a certain attention to their privacy.
- **RQ**$_2$: *What types of private information are disclosed on GitHub by users?* Although designed as a platform for technical purposes, GitHub inherently possesses social media characteristics. Consequently, our study investigates the information that developers disclose, whether intentionally or unintentionally, in their pull_requests comments.
- **RQ**$_3$: *After users choose their privacy settings, do they adhere to what they have declared? In other words, can we observe a discrepancy between their stated privacy preferences and their actual behavior, such as their textual activity?* It is relevant to understand whether there is a mismatch to assess the effectiveness of these privacy settings and to facilitate the development of personalized privacy assistant for users.
- **RQ**$_4$: *To which extent is it possible to automate the detection of sensitive comments with the use of BERT or Llama2?* We explored if the sensitivity detection of textual comments on this platform can be conducted by leveraging models as BERT or Llama2.

The main contributions of our work are summarized below:

- We conducted an empirical study on a set of 6,132 developers to investigate the privacy dynamics in the GitHub ecosystem.
- Our investigation reveals that users adopt different configurations of privacy settings, showing different privacy concerns.
- The empirical study shows that users reveal different types of information about themselves or other developers through pull_request comments. This triggers the need to overhaul the privacy management from the designers of GitHub.
- Our ultimate goal is to encourage the development of innovative applications for detecting privacy data leakage. Specifically, we present a proof of concept where we train BERT and we fine-tuned Llama–a large language model (Touvron et al. 2023) (LLM)–to automatically assess whether a text discloses sensitive information.
- The dataset curated through this paper has been published[2] to facilitate future research (see Section 8).

[2] https://github.com/MDEGroup/EMSE-CHASE-Privacy

The paper is structured as follows. In Section 2, we provide a background on privacy settings on GitHub, a set of vulnerabilities that show the need for techniques and tools to manage privacy settings in the platform, and a real case scenario. Section 3 reviews related work. Section 4 presents the methodology to perform the empirical study on the GitHub ecosystem. Section 5 presents the proof of concepts with BERT and Llama2. The results are reported and analyzed in Section 6. We discuss the results in Section 7. Section 8 sketches future work, and concludes the paper.

2 Motivations and Background

In this section, we illustrate the different motivations behind this study and provide background on privacy settings on GitHub and on GitHub privacy vulnerabilities.

2.1 GitHub and its Privacy Settings

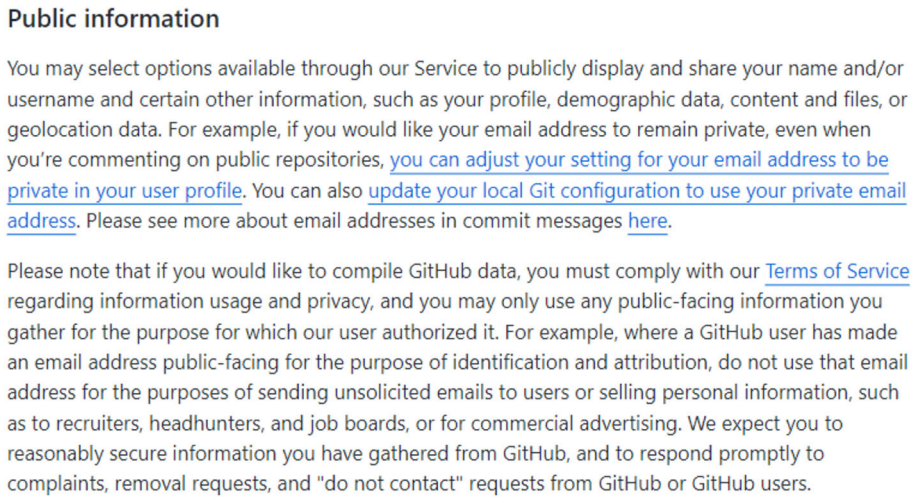
We chose to investigate GitHub for different reasons: *(i)* GitHub is a platform for technical purposes and may appear neutral or devoid of any privacy or ethical concerns. However, "[open source software] is as much social as it is technical" (Vasilescu et al. 2015c) and there is already evidence from the literature that proves privacy infringements and gender discrimination on the platform (Terrell et al. 2017; Meli et al. 2019; Ford et al. 2019; Niu et al. 2023); *(ii)* On this platform, there is a large availability of data that can be downloaded as well as dataset already processed (Gousios 2013); *(iii)* Furthermore, this platform allows us to access users' privacy settings thus permitting to compare users' stated privacy preferences with their actual behaviors. This approach represents a novelty compared to previous studies on the utilization of privacy settings on platforms, which relied solely on user surveys rather than an analysis of actual preference selections (Kanampiu and Anwar 2019; Chen et al. 2019). The analysis of privacy preferences offers a deeper understanding of user behaviors and the limitations associated with the available privacy settings.

In this study, we concentrated on examining users' selection of privacy preferences and their self-disclosure of personal information in pull_request comments. Our objective is to analyze the privacy dynamics on GitHub, i.e., whether users effectively used the privacy settings provided, if these settings present any limitations, and if any personal information, intentionally concealed or not, can be inferred from a user's textual activity on GitHub. Textual activities vary from making a commit, sending a pull_request, commenting on all these different activities. In this study, we focused on the task of commenting on a pull_request. This decision was guided by previous studies highlighting a greater probability of encountering significant user interactions and consequently more sensitive information in pull_request contexts (Sajadi et al. 2023; Iyer et al. 2019).

According to the GitHub Privacy Statement (see also Fig. 1),[3] a user who does not wish to show all the information available on her GitHub profile, can adjust the privacy settings provided by the platform in order to hide pieces of information.

When referring to users' privacy, we talk about *personal* users' privacy, i.e., what they aim to share about their life. On GitHub, these *desiderata* can be expressed through the privacy settings of the profile. An example of what can be shown or hidden on GitHub is illustrated in Fig. 2. We consider the totality of the privacy settings selected by the users as their *privacy desiderata*.

[3] https://bit.ly/461tL9i



Public information

You may select options available through our Service to publicly display and share your name and/or username and certain other information, such as your profile, demographic data, content and files, or geolocation data. For example, if you would like your email address to remain private, even when you're commenting on public repositories, you can adjust your setting for your email address to be private in your user profile. You can also update your local Git configuration to use your private email address. Please see more about email addresses in commit messages here.

Please note that if you would like to compile GitHub data, you must comply with our Terms of Service regarding information usage and privacy, and you may only use any public-facing information you gather for the purpose for which our user authorized it. For example, where a GitHub user has made an email address public-facing for the purpose of identification and attribution, do not use that email address for the purposes of sending unsolicited emails to users or selling personal information, such as to recruiters, headhunters, and job boards, or for commercial advertising. We expect you to reasonably secure information you have gathered from GitHub, and to respond promptly to complaints, removal requests, and "do not contact" requests from GitHub or GitHub users.

Fig. 1 GitHub Privacy Statement on settings

2.2 Privacy Vulnerability in GitHub

Privacy vulnerability arises in several places. Personal information such as email addresses, location, and potentially even real names are often part of a user's profile. Even if GitHub provides tailored settings for privacy (see Fig. 2), the activities of users, such as the repositories they star, the issues they comment on, or the projects they follow, can reveal a lot about their interests, expertise, and professional activities. Due to the social nature of GitHub, user collaborations can be used to identify user's professional network, which might include sensitive information, especially for those working on confidential or competitive projects.

Privacy vulnerabilities can arise as follows:

V1 The activities of a user who has chosen to set her event settings to private can potentially be discovered by examining the history of repositories to which she has made contributions.

V2 When users make links that expose personal or sensitive information, their *privacy desiderata* may be compromised.

V3 Public visibility occurs when users fork or star each other's repositories. These actions provide information about users' hobbies, projects, or collaborations. Additionally, issue and pull request discussions are public by default, revealing complex collaboration net-

(a) Personal information (b) Contribution settings on GitHub

Fig. 2 Privacy settings on GitHub

works. A user frequently addressing user-initiated issues or pull requests, may suggest a relationship beyond GitHub, such as coworkers or acquaintances. Many organizations and teams are open to the public. This could reveal confidential professional relationships or collaborations. The public exposure of the user's contribution graph, which shows their work patterns and hours, might reveal work habits and inactivity. This information may be linked to job absence or personal activities.

V4 Textual contents such as commit messages, Pull requests, and issue comments may disclose privacy information.

The above-listed privacy vulnerabilities are well-known to researchers who publish under double-blind peer review contexts (Bacchelli and Beller 2017). In many research environments, double-blind peer review is pivotal to the integrity of knowledge dissemination. This requires to ensure the anonymity of submission materials, including associated replication packages. On GitHub, users need to anonymize sensitive information to protect privacy while still allowing for the replication of their results. Although GitHub allows users to specify privacy settings for their repositories reducing the likelihood of disclosing user identities, it is necessary to thoroughly analyze repository content, metadata, and textual comments (e.g., commit, issue, and pull request comments) to delete any trace that can reveal identities or affiliations. In this context, external tools are developed to remove user data from repositories, e.g., *Anonymous GitHub*.[4] However, the risk of revealing the user identity still exists as those tools mainly apply rewriting rules, where the user should provide the complete list of terms that will be anonymized. Moreover, to the best of our knowledge, no tool is provided for sanitizing textual comments which represent a threat to privacy, as we show in this paper. Ideally, any automated means should take in input the user's privacy desiderata and then act to support the user in maintaining her privacy desiderata throughout her interactions with the system.

2.3 The Recruiters Problem

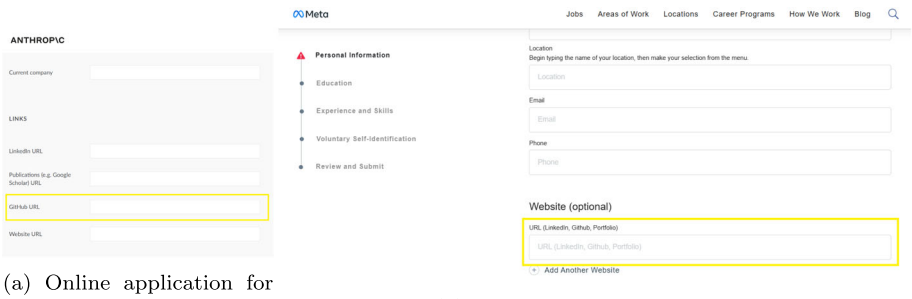
It is widely acknowledged that recruiters leverage social networks to gather insights into their candidates' backgrounds (Becton et al. 2019; El Ouirdi et al. 2016). The information sought varies, including personality traits, communication skills, the presence of provocative or inappropriate photographs, or any other factors that might dissuade the hiring of a candidate (Henderson 2019).

A survey conducted by CareerBuilder[5] reveals that 70% of employers utilize social media as a screening tool for potential hires. Despite the potential utility or innocuous intentions perceived by recruiters, this practice raises concerns about potential discrimination against candidates and the disclosure of information forbidden during the interview process. For example, Acquisti and Fong (2020) showed how the disclosure of certain personal information online can influence the hiring decisions of some U.S. employers, confirming the problematic nature of this practice.

Currently, it is a common practice for companies to request a candidate's GitHub profile as part of the hiring process. Figure 3 displays a screenshot of the online application procedures for two renowned companies: Anthropic (Fig. 3a) and Meta (Fig. 3b). In addition to the

[4] https://anonymous.4open.science/

[5] https://prn.to/3NugUWa



(a) Online application for Anthropic.

(b) Online application for Meta.

Fig. 3 Examples of companies requiring the GitHub profile

conventional inclusion of LinkedIn profiles and links to Google Scholar publications, both companies demand the inclusion of the GitHub URL.

The companies under consideration are technology giants, and requesting the GitHub profile could constitute a component of the verification process for assessing technical skills. However, on GitHub, numerous interactions among users have been scrutinized by various researchers to tackle and alleviate social issues such as toxicity or stress (Miller et al. 2022a; Raman et al. 2020), showing that the usage of this platform goes beyond the sharing of technical knowledge.

Understanding the nature of information accessible on GitHub is not only essential for the evidence we gathered regarding recruiter interest in this platform, but also from a broader perspective as it contributes to the overall "digital breadcrumbs" left by users (Stretton and Aaron 2015). The significance lies not solely in the context of a single platform; rather, it pertains to the aggregation of information within the digital realm that can be associated with a user and exploited for different purposes, such as surveillance or psychological targeting (Lustgarten et al. 2020; Matz et al. 2020; Miller 2010).

3 Related Work

This section reviews empirical studies on GitHub and findings on privacy issues on this platform. Moreover, we review works on the analysis of user-generated content and users' privacy. We focus on works that have produced labeled corpora as well as privacy profiling.

3.1 GitHub Studies

GitHub is a widely used platform, and numerous researchers have conducted empirical studies on this platform. In their exploratory study, Henning et al. (2023) examined reported issues related to data protection on GitHub, shedding light on the influence of data protection regulations throughout the entire software development process. Acar et al. (2017) conducted a study aimed at enhancing security decision-making for IT professionals. The researchers conducted an experiment with 307 active GitHub users to assess their performance on security-related programming tasks. They found differences in performance related to experience levels, but no significant disparities based on student or professional status. Khalajzadeh et al. (2022a) conducted an analysis of comments on GitHub in order to uncover issues that are centred around human concerns. The analysis revealed a diverse array of issues, encompassing topics such as privacy and security.

Many researchers have focused on investigating a more social aspect of GitHub. Sajadi et al. (2023) explored the dimension of interpersonal trust in Open Source Software (OSS) teams and how it is exhibited. The study analyzes 100 GitHub pull requests from Apache Software Foundation projects to understand how trust is expressed in these interactions. Guzman et al. (2014) investigated the impact of emotions on productivity, task quality, creativity, group relationship, and job satisfaction through sentiment analysis of commit comments in various open-source projects. The results indicate that Java projects tend to receive more negative commit comments, while projects with widely distributed teams tend to have more positive emotional content. Raman et al. (2020) studied the problem of stress and burnout in open-source environments due to toxic discussions on GitHub issues. They demonstrated that a combination of pre-trained detectors for negative sentiment can effectively identify these issues. Furthermore, they established that classification accuracy is enhanced through domain adaptation. In their study, Blincoe et al. (2016) defined popular users as individuals who provide guidance to OSS developers when they join new projects. The authors observed that those users did not possess the highest contribution rate. Miller et al. (2022b) curated a sample of 100 toxic GitHub issue discussions to gain an understanding of the characteristics of open-source toxicity. They found that some of the most prevalent forms of toxicity are entitled, demanding, and arrogant comments from project users as well as insults arising from technical disagreements.

Different authors have investigated more specifically solely the topic of privacy leakages and self-disclosure on GitHub. According to Vasilescu et al. (2015c), their user survey revealed that platform users are aware of certain demographic details about other developers, including gender, real names, and countries of residence. Additionally, Ford et al. (2019) demonstrated that GitHub developers explore a much wider array of information while scrutinizing pull requests, particularly information tied to the identity of the person submitting the pull request. This implies that GitHub not only contains identity information but also that such information is exploited during the evaluation of pull requests. In a similar vein, Meli et al. (2019) discovered that hundreds of thousands of API and cryptographic keys are leaked on GitHub at a rate of thousands per day. Meanwhile, Niu et al. (2023) demonstrated that it is possible to extract sensitive personal information from the Codex model used in GitHub Copilot.

From these studies, it is evident that GitHub has garnered attention from the software engineering community for several years. The primary findings indicate that various user information, including gender, real names, and countries, can be extracted from GitHub. Furthermore, the platform faces general privacy threats, such as the leakage of crypto-related secrets and private information on Copilot. Additionally, GitHub exhibits characteristics akin to other social networks, such as language toxicity, and is subject to study in the context of team working and cooperation. It is worth noting that none of the mentioned studies investigated the adoption of privacy settings provided by the platform. Previous research has explored the adoption of privacy settings on other platforms such as Facebook (Fiesler et al. 2017; Chen et al. 2019; Kanampiu and Anwar 2019), primarily through user surveys. To the best of our knowledge, this is the first study on GitHub privacy settings that is conducted based on the actual selections made by users.

3.2 Analysis of User-Generated Content

User-generated content plays a significant and ubiquitous role on various platforms, attracting extensive attention from researchers for diverse purposes such as sentiment analysis, risk

detection of depression (e.g., on Reddit), marketing analysis, and self-disclosure (Alaei et al. 2023; Tadesse et al. 2019; Timoshenko and Hauser 2019; Umar et al. 2019). In this section, we present an overview of studies focused on the analysis of user-generated content, with a specific emphasis on the detection of self-disclosure.

Bioglio and Pensa (2022) conducted a study aimed at automatically detecting the sensitivity of Facebook posts. Their research involved analyzing a dataset comprising 9,917 Facebook posts, each one annotated by three experts as either sensitive, non-sensitive, or of unknown sensitivity. To enhance their investigation, two additional datasets were incorporated for comparative analysis. The first dataset consisted of posts extracted from Reddit, manually labeled to align with the Facebook corpus. The second dataset comprised anonymous posts from Whisper, considered sensitive, alongside non-sensitive tweets sourced from Twitter (now renamed as **X**).

In their experiment, the researchers employed four distinct classifiers on the datasets: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) with gated recurrent unit, RNN with long short-term memory, and BERT. Notably, they highlighted the limitations of existing corpora for sensitivity detection, emphasizing that many are often derived from specific topics, and thus incapable of detecting sensitivity on wider topics. Furthermore, their findings demonstrated that, on their datasets, RNNs and BERT exhibited significant performance in classification, and that lexical features are not sufficient for discriminating between sensitive and non-sensitive texts.

In their examination of social media safety, DiSalvo et al. (2022) introduced a methodology for categorizing social media posts by utilizing a pre-established corpus of violent phrases. Specifically focusing on Twitter, the researchers collected posts that were then organized using the labelTweets function. This function evaluates a tweet based on whether it contains a word from a predefined array of 'violent' terms, either accepting or rejecting the tweet accordingly. Subsequently, the authors manually assigned labels (negative, positive, or neutral) to the output generated by the labelTweets function, guided the corresponding corpora. The resulting corpus comprised nearly 600 tweets. The authors proceeded with a classification task to differentiate between negative and positive tweets, employing three classifiers: Naive Bayes Classifier, Support Vector Machine, and Logistic Regression. Their analysis concluded that the Naive Bayes Classifier performed the least effectively among the three.

Blose et al. (2020) conducted a study on self-disclosure patterns on social media, specifically focusing on tweets posted during the Coronavirus pandemic. The authors utilized a dataset comprising Tweet IDs, which was updated through the Amazon Web Service. Their methodology involved a pre-processing of the dataset, for filtering only content posted by individual users. To automate the detection of self-disclosure, the researchers employed a dictionary obtained from Umar et al. (2019) and compared their dataset against an already annotated Twitter dataset. This comparison yielded satisfactory results, demonstrating the effectiveness of their approach. To better understand disclosure trends, the authors conducted a topic-modeling analysis on the described corpus. Furthermore, they conducted a comparative analysis, juxtaposing their observations of self-disclosure behaviors during the Coronavirus pandemic with those observed during Hurricane Harvey in 2017. The findings of the study revealed an increase in self-disclosure behaviors during the pandemic. In their conclusion, the authors encourage further research engagement to "better understand more subtle, voluntary self-privacy violations." This underscores the importance of ongoing investigations into evolving patterns of online self-disclosure, especially during critical events such as the Coronavirus pandemic.

Despite their distinct objectives, all these studies analyzing user-generated content share a common procedure, which can be summarized as follows: selecting a specific platform or online social network, retrieving textual data from this platform for preprocessing, and utilizing domain-specific resources like the Privacy Dictionary (Gill et al. 2011; Vasalou et al. 2011) or the self-disclosure dictionary (Umar et al. 2019) to support the preprocessing phase. Ultimately, classifiers are employed to automate the detection process.

In our research, we adhered to the same methodology described above for self-disclosure detection in pull_requests comments. We focused on creating a multi-label corpus, where each comment is annotated for every sensitive information shared in the same text.

3.3 User's Privacy Studies

With the increasing attention to the issue of privacy in the digital world, different researchers have investigated how to capture users' privacy desiderata and users' behaviours in online platforms.

Brandão et al. (2022) analyzed users' privacy profiles for what concerns mobile settings and exploited this information to predict the users' answers to a permission request, while preserving user privacy by applying a federated learning approach.

Tahaei et al. (2020) examined Stack Overflow for privacy-related challenges faced by developers. They used topic modelling on 1,733 questions, identifying key topics. The results indicated that developers do seek support on privacy issues on Stack Overflow.

In an attempt to study privacy issues in Twitter (now **X**) (Keküllüoglu et al. 2020), the authors collected a dataset of 635k tweets containing the expression *"happy for you."* By employing LDA topic modeling, the tweets were categorized into 12 distinct clusters representing different life events. They found out that around 8% of the tweets mentioned protected users, with varying rates across different topics.

4 Methodology

A primary goal of this study is to investigate the use of privacy settings on GitHub, that is, users' privacy desiderata expressed on the platform, and their potential limitations (RQ1). To this end, we started by exploring the largest existing dataset of GitHub activities, the GHTorrent dataset (Gousios 2013; Gill et al. 2011). In this dataset, the *Users* table provides the set of privacy settings selected by each GitHub user as their privacy preferences on the platform. We adopted this information as being users' privacy desiderata. Furthermore, from the GHTorrent dataset, we defined and selected a subset of users being more "active" on GitHub, in order to conduct a more fine-grained analysis and to have more textual data to analyze. We called this dataset the *Active users*. We updated the *Active users* dataset due to the absence of certain privacy settings that were missing from the original GHTorrent dataset, such as email addresses and social media channels. This process is detailed in Section 4.1. By strategically retrieving additional data from the current database, it was possible to effectively address the gaps in users' metadata and ensure a thorough understanding of user data. This was achieved through the use of GitHub APIs[6] and the users' login information, which allowed for the seamless integration and enrichment of GHTorrent existing data with detailed and updated user-specific insights. On both datasets (*Users* and *Active users*), we conducted a cluster analysis to define users' privacy profiles and observe differences between users

[6] https://docs.github.com/en/rest?apiVersion=2022-11-28

selection of settings. All these steps are detailed in Sections 4.1, 4.2, and represented in Fig. 4.

To investigate RQ2 and RQ3, that is, the information disclosed on the platform and the analysis of users' behavior, we considered textual comments provided by the GHTorrent dataset. In particular, we analyzed the pull_request comments to find privacy-sensitive information disclosed intentionally or unintentionally by the users. The methodology for this analysis is described in Section 4.3. In addition, it should be noted that GHTorrent truncates textual comments, specifically those found in commits, issues, and pull_requests. In order to facilitate the empirical investigation and analysis of privacy awareness in pull_requests, we acquired the original text through the use of GitHub APIs.

4.1 Data Curation

When talking about users' private information, we refer to the personal information they do not want to share. On GitHub, these *privacy desiderata* can be expressed through the profile settings where the user chooses to reveal or hide certain information (company, location, bio, social accounts, and so on). In the GHTorrent dataset (Gousios 2013; Gill et al. 2011), this information was collected in the *User's* table (see Table 1). For each user in the GHTorrent dataset, we were able to retrieve information regarding whether the user chose to share the name of their company on their profile, along with details about their location, including city and state. These choices reflect the decisions made by users at the time of the last GHTorrent updates, which occurred in 2019. To overcome these limitations, we updated the privacy choices of the *Active users* dataset. The data curation process is summarized in Table 2.

Step 1: GHTorrent preprocessing Users We used the GHTorrent dataset to conduct a cluster analysis on a broader population of GitHub users.

The *Users'* table from the GHTorrent dataset contains 32,411,628 accounts, from which we eliminated those marked as `organization`, `fake` and `deleted`. This resulted in a set of 22,525,012 users. Figure 5 shows the correlation matrix of the variables in *Users* table, obtained using the Python library Scikit-learn. Based on this correlation matrix, we excluded the `state`, `country code`, `lat`, and `location` variables because of the value of correlation being very high (above 0.7), following the methodology outlined by King (2015).

Every column of the dataset was converted into 0 if the value was missing, and 1 if the value was present. This choice was made because we were only interested in evaluating whether a user would disclose (1) or not (0) that piece of information. The resulting dataset is

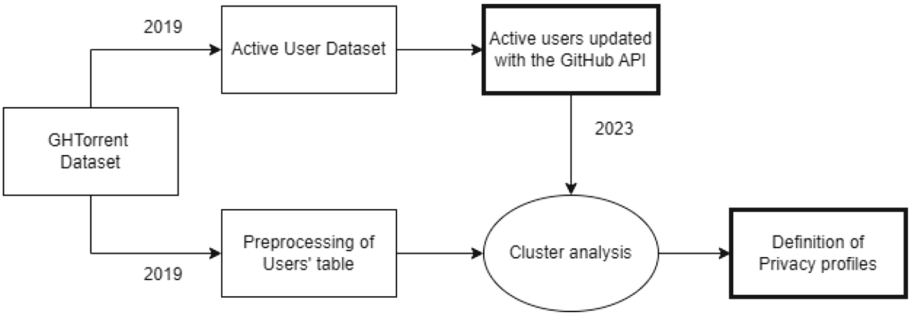


Fig. 4 Workflow of the study

Table 1 Table *Users* from the GHTorrent dataset

login	company	created_at	type	fake	deleted	long	lat	country code	state	city	location
U_1 [redacted]	–	2016-04-18 11:42:46	USR	0	1	–	–	–	–	–	–
U_2 [redacted]	Sage GmbH	2008-12-15 12:28:33	USR	0	0	0.00000000	0.00000000	–	–	–	Rastede, Germany
U_3 [redacted]	–	2008-03-22 00:37:42	USR	0	0	132.45529270	34.38520290	jp	Hiroshima Prefecture	Hiroshima	Hiroshima
U_4 [redacted]	–	2012-08-03 16:08:15	USR	1	1	–	–	–	–	–	–
U_5 [redacted]	@ValuationUp	2008-04-28 17:25:53	USR	0	0	28.04730510	-26.20410280	za	City of Johannesburg	Johannesburg	Johannesburg, South Africa
...	...	...	...	...	...	...	...	...	...	...	...

Table 2 Data curation process

Step	Source	#Users	#Pr Comments
Step 1	GHTorrent	22,525,012	38,159,840
Step 2	*Active users*	6,329	89,749
Step 3	GitHub API actualization	6,132	15,672

represented in Table 3. In the tables, we hide information related to user identification using the [redacted] field, due to privacy reasons.

Step 2: Active users To have a more updated version of users' privacy preferences, we restricted our research to those users considered more "active" on GitHub, those who may be identified by quantifying the number of actions they have performed on the platform. Therefore, we constructed a new dataset that contained information about the number of `commits`, `commit_comments`, `followers`, `pull_requests` `comments` and `issue_comments` executed by each user (see Table 4). This achievement was made possible through the GHTorrent dataset since this dataset allows the extraction of the precise count for each action executed by the users. Indeed, we retrieved this information from the corresponding tables from the GHTorrent dataset. After the due preprocessing step, we performed a cluster analysis on the dataset described in Table 4. Primarily, we scaled this dataset, and performed a K-means cluster analysis on the dataset with a number of clusters K=4 (found with the Elbow method).

Secondly, we discretized the variables into three distinct bins. The "low" label was assigned to instances where the number of actions ranged from the minimum value to the 65th percentile. The "medium" label was applied when the number of actions fell between the 65th

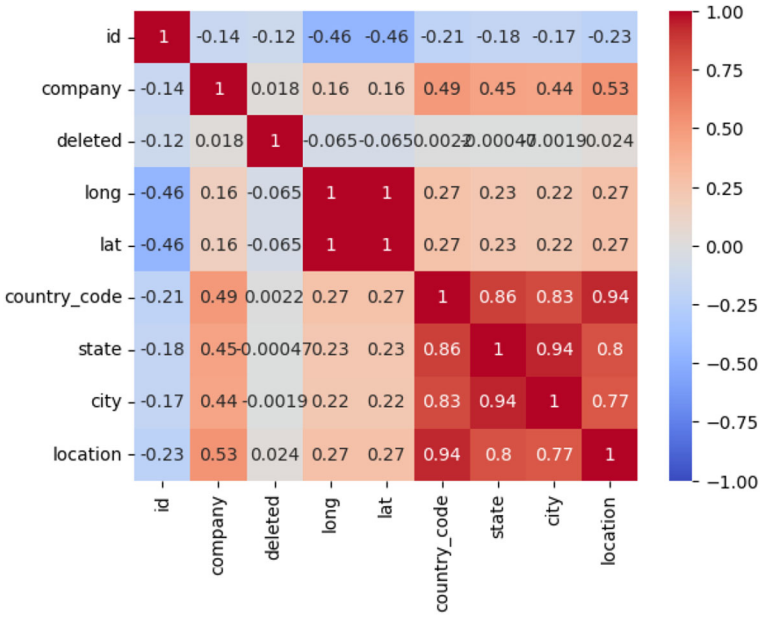


Fig. 5 Correlation matrix of the variables in the *Users* dataset

Table 3 *Users*' table preprocessed

login	company	long	city
U_1 [redacted]	1	1	0
U_2 [redacted]	0	1	1
U_3 [redacted]	0	0	0
U_4 [redacted]	1	1	1
U_5 [redacted]	1	1	1
...	...	...	...

percentile and the mean value. Finally, the "high" label was assigned to cases where the number of actions varied from the mean value to the maximum value. The choice of the 65th percentile as a threshold was deliberate, as it consistently yielded values lower than the mean across all variables in the dataset, as evidenced by the statistics in Table 5.

Figure 6a shows the different clusters of users, according to the number of actions performed. Figure 6b, c, d, and e illustrate the percentage of each variable per cluster. For instance, examining Fig. 6b reveals that users in Cluster 0 tend to make few commit comments (the variable 'low' is predominant in this cluster). The analysis of the variables per cluster in Fig. 6 allowed us to clearly identify the *least* active users, being cluster 0. Consequently, we excluded this cluster of *Non-Active users* from the overall dataset, resulting in what we refer to as the *Active users* dataset (6,329 users in total). The number of *Active users* decreased to 6,132 with the updates of users' privacy settings that we have performed through the GitHub API, as explained in the next step. This is attributed to users departing the platform by 2023.

Step 3: Updating active users privacy settings Once we selected the *Active users*, we updated their privacy settings choices using the GitHub API with their login. This action was taken due to the incomplete nature of the GHTorrent dataset, which lacked information such as users displaying their email addresses, events, and their Twitter accounts. The updated dataset is illustrated in Table 6. Compared with the GHTorrent dataset, the columns added were `email`, `events` and `twitter`. Some of the users in the *Active users* dataset were no longer present on GitHub at the time of the update so they were eliminated. The final number of updated *Active users* is 6,132 users. The methodology detailed in this section is represented in Fig. 4.

After the data curation described in Section 4.1, we conducted a cluster analysis on the two datasets: *Users* and *Active users*. The goal was to understand whether users present variability in terms of privacy settings on GitHub, i.e., if there are different *privacy profiles* that can be

Table 4 Dataset of actions performed by each user on GitHub

user_id	#pull request comments	#followers	#commits	#commit comments	#issue comments
0	11.0	–	–	14.0	–
1	15.0	31.0	21.0	6.0	30.0
2	25.0	400.0	3598.0	25.0	1504.0
3	4.0	–	2.0	3.0	–
4	5.0	174.0	463.0	14.0	516.0
...	...	...	...	...	...

On this dataset, we performed a cluster analysis to define the more "active" users on the platform

Table 5 Statistics from the dataset of actions performed by users on GitHub

	#pull request comments	#followers	#commits	#commit comments	#issue comments
Count	1.2×10^{5}	1.2×10^{5}	1.2×10^{5}	1.2×10^{5}	1.2×10^{5}
Mean	4.9×10^{-4}	1.5×10^{-3}	3.9×10^{-3}	2.5×10^{-3}	4.9×10^{-3}
Std	4.4×10^{-3}	9.7×10^{-3}	1.2×10^{-2}	1.2×10^{-2}	1.5×10^{-2}
Min	0.0×10^{0}	0.0×10^{0}	0.0×10^{0}	0.0×10^{0}	0.0×10^{0}
25%	0.0×10^{0}	1.3×10^{-4}	7.9×10^{-5}	2.5×10^{-4}	3.0×10^{-4}
50%	0.0×10^{0}	3.7×10^{-4}	7.1×10^{-4}	5.1×10^{-4}	1.0×10^{-3}
65%	3.3×10^{-5}	6.1×10^{-4}	1.8×10^{-3}	1.0×10^{-3}	2.1×10^{-3}
75%	9.8×10^{-5}	9.3×10^{-4}	3.3×10^{-3}	1.7×10^{-3}	3.5×10^{-3}
Max	1.0×10^{0}	1.0×10^{0}	1.0×10^{0}	1.0×10^{0}	1.0×10^{0}

associated with the users of this platform (Di Ruscio et al. 2024) meaning that users adopt different privacy settings as a tool for their privacy. The clustering techniques adopted were K-means and hierarchical clustering, as suggested by different authors (Sanchez et al. 2020; Brandäo et al. 2022).

4.2 Cluster Analysis

To define the privacy profiles of GitHub users, we performed cluster analysis on the two datasets, *Users* and *Active users*. The variables considered are the privacy settings chosen by each user, however, there are variations between the two sets due to recent additions of privacy features by GitHub, such as `events` and `twitter`. This novelty is reflected in the updated dataset of *Active users*. In the *Active users* dataset, we have added the `email` field, which was removed from the *Users* dataset by the author of the GHTorrent dataset.

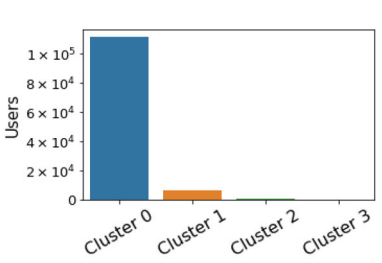
For the *Users* dataset preprocessed as described in Section 4.1 *Step 1*, we performed a K-means cluster analysis (Hartigan and Wong 1979) with Euclidean distance. The number of clusters K=3 was chosen with the Elbow method (Syakur et al. 2018).

For the *Active users* dataset updated as described in Section 4.1 *Step 3*, we removed the variables `location`, `long`, `lat`, `country code`, and `state` from the *Active users* dataset due to their correlations as shown by the correlation matrix in Fig. 7. We employed both the Elbow method (Fig. 8) and an analysis of the dendrogram obtained using Ward's method (Fig. 9) to determine an appropriate number of clusters. We identified K=4 as a reasonable number of clusters for this dataset, therefore a K-means cluster analysis was performed with this value.

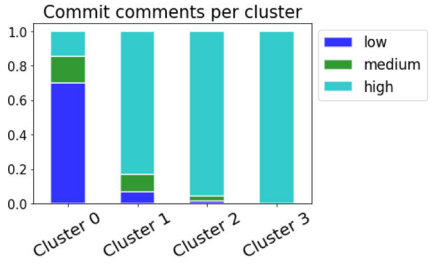
Both cluster analyses for the datasets *Users* and *Active users* were conducted in order to study users' privacy profiles and to verify whether these settings were actively adopted by the users. The results are discussed in Section 6.

4.3 Construction of the Corpus

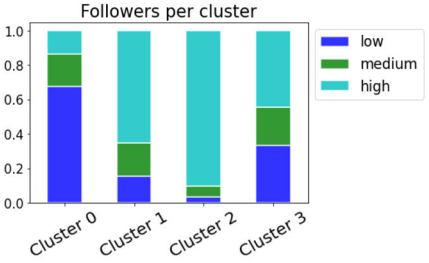
To address RQ2 and RQ3, we started exploring users' privacy behaviour for what concerns textual data, e.g., their comments on GitHub. While there's a wealth of textual data in the GHTorrent dataset, we focused our research on pull_requests comments exclusively (86,000



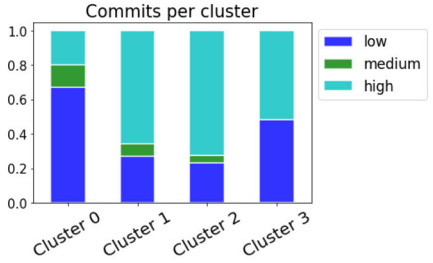
(a) Users per cluster



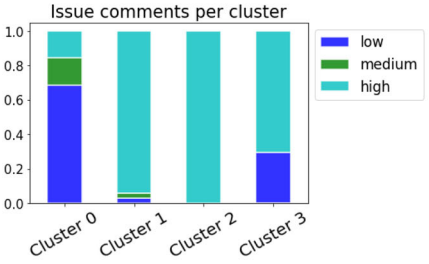
(b) Commit comments



(c) Followers



(d) Commits



(e) Issue comments

(f) Pull requests comments

Fig. 6 Cluster analysis on the dataset of the actions performed by each user. Figure 6a shows the number of users per cluster, Fig. 6b, c, d, and e show the distribution of variables per cluster

comments overall).This decision was guided by previous studies highlighting a greater probability of encountering significant user interactions and consequently more sensitive information in pull_request contexts (Sajadi et al. 2023; Iyer et al. 2019).To collect the more privacy-sensitive data and prepare the dataset for subsequent manual labeling, we automatically labeled each comment using the Privacy Dictionary created by Vasalou et al. (2011); Gill et al. (2011), exploiting libraries provided by existing work (Casillo et al. 2022).This dictionary was constructed and validated by its authors through interviews and focus groups from different privacy-sensitive (offline and online) contexts, leading to identify eight different privacy categories, as illustrated in Table 7. Each category represents a distinct privacy realm, potentially encompassing different types of private information. Previous authors have successfully adopted this dictionary to detect privacy language patterns within a given text, such as Bioglio and Pensa (2022) and D'Acunto et al. (2021). The final goal of this process

Table 6 The *Active users* dataset updated through the GitHub API

login	company	created_at	location	long	lat	country code	state	city	email	events	Twitter
U_1 [redacted]	@floraison	2008-03-22 00:37:42	Hiroshima	132.45529270	34.38520290	jp	Hiroshima Prefecture	Hiroshima	E_1 [redacted]	151	–
U_2 [redacted]	DNSimple	2008-04-06 08:44:35	Rome, Italy	12.49636550	41.90278350	it	Rome	Rome	E_2 [redacted]	161	T_2 [redacted]
U_3 [redacted]	–	2010-02-05 06:35:04	–	0.00000000	0.00000000	–	–	–	–	30	–
U_4 [redacted]	@platformatic	2009-02-05 21:24:19	Forlì, Italy	0.00000000	0.00000000	–	–	–	E_4 [redacted]	291	T_4 [redacted]
U_5 [redacted]	–	2011-01-23 18:48:07	–	0.00000000	0.00000000	–	–	–	E_5 [redacted]	284	–
U_6 [redacted]	Cocos	2010-11-19 08:06:45	Xiamen Fujian China	118.08942500	24.47983300	cn	Fujian	Xiamen	E_6 [redacted]	290	T_6 [redacted]

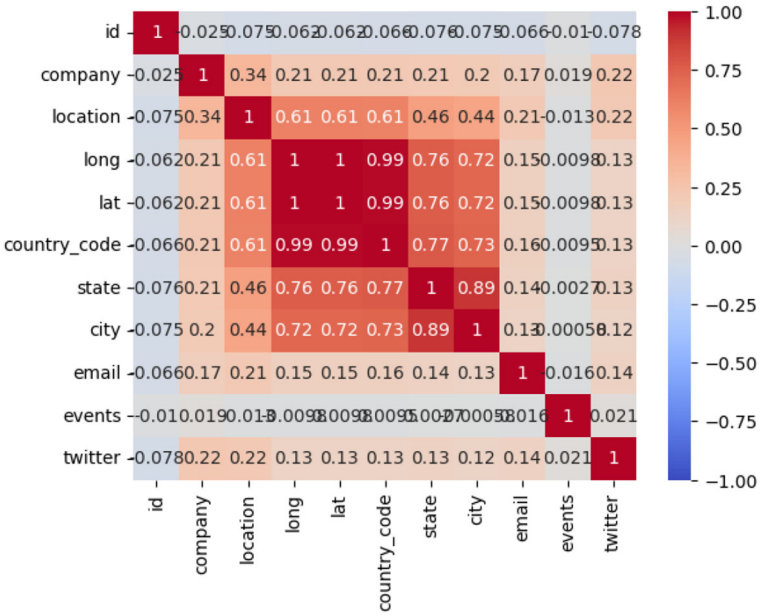


Fig. 7 Correlation matrix of the variables present in the *Active users* dataset

is to gather evidence of self-disclosure on GitHub. An example from the corpus labeled with the Privacy Dictionary is shown in Table 8. The first column indicates the user who made the comment in the "body" column. The "Categories" column shows the privacy category assigned to each comment, identified through the "Keywords" in the corresponding column.

Given that GitHub pull_request comments often involve technical details, we aimed to enhance the efficiency of identifying comments with private information. To achieve this,

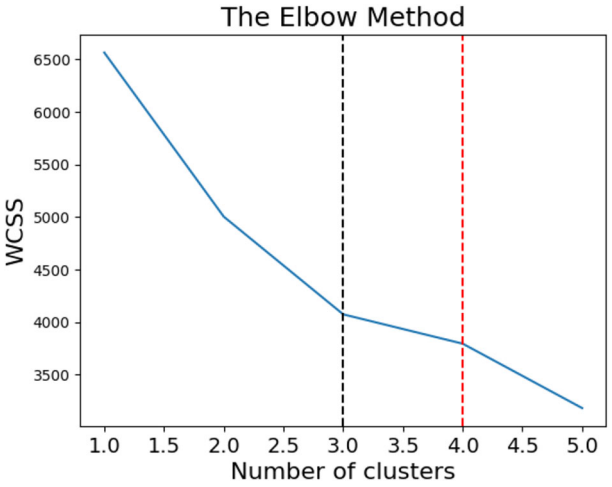


Fig. 8 Elbow method on the *Active users* dataset to establish the more appropriate number of clusters

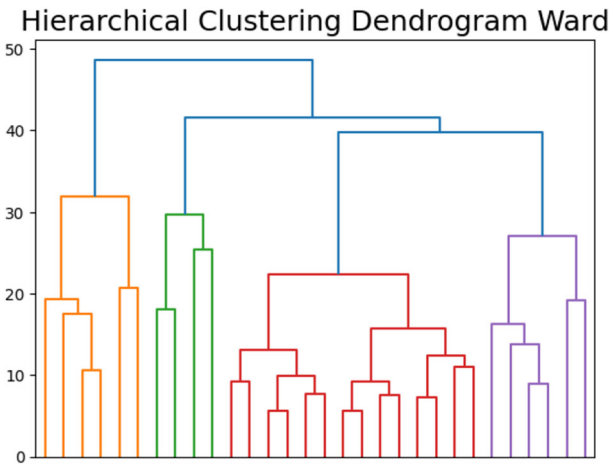


Fig. 9 Dendrogram with Ward's method on the *Active users* dataset to have an overview of the dataset and to confirm the appropriate number of clusters

we specifically focus on comments that carry more than one label from the categories within the Privacy Dictionary. Indeed, due to the broad scope of the dictionary, we have empirically noted that most comments were assigned at least one label. Therefore, we chose to select comments with multiple labels that potentially included information from various privacy

Table 7 The eight privacy categories as described in the Privacy Dictionary

Category	Definition
NegativePrivacy	This category captures the antecedents and consequences of breaches in privacy. It encompasses terms associated with privacy concerns and risks as well as judgments about the source and type of violation.
NormsRequisites	NormsRequisites includes the norms, beliefs, and expectations in relation to achieving privacy.
OutcomeState	It includes words that describe the static behavioral states and the outcomes that are served through privacy. This grouping corresponds to Westin's (1967) delineation of privacy states and purposes.
PrivateSecret	It includes descriptors or terms that articulate the essence of privacy. This category aids in discerning precisely which elements individuals perceive as private.
Intimacy	Intimacy comprises words that portray and measure different facets of small-group privacy. It includes words that denote the psychological needs involved in revealing oneself to another person, as well as the emotional proximity that forms between individuals.
Law	This category includes words employed to describe legal definitions of privacy.
Restriction	Words in this category express the closed, restrictive, and regulatory behaviors employed in maintaining privacy. Thus, the Restriction category can be used to measure the behaviors that people take to protect their privacy.
OpenVisible	This category includes words that represent the dialectic openness of privacy.

Table 8 Corpus labeled using the Privacy Dictionary

user_id	body	login	Categories	Keywords
2	Hello Jun, I would prefer if the conversation ...	U_1 [redacted]	Intimacy OpenVisible	['conversation']
2	Could you please open an issue for that?Thanks...	U_1 [redacted]	OpenVisible	['open']
5	Makes sense. I'll hold off this change to a se...	U_2 [redacted]	OutcomeState	['separate']
5	FYI, I started a prototype long time ago:https...	U_2 [redacted]	PrivateSecret	['identity']
10	ahaha! I think it was a test that it did not w...	U_4 [redacted]	OutcomeState Restriction	['safely', 'delete']
10	this whole block on code should be indented of...	U_4 [redacted]	Restriction	['block']
10	sure! I'm kind of full at the moment, but I'll...	U_4 [redacted]	OpenVisible	['report']
10	This is not completely correct, as calling 'th...	U_4 [redacted]	OutcomeState	['prevent']
10	Can you please release this? It's completely o...	U_4 [redacted]	OpenVisible OutcomeState	['release']
10	no idea. We have that code running in aws-ami-...	U_4 [redacted]	OutcomeState	['safe']
...	...	...	...	...

domains represented by each privacy category of the dictionary. This approach increases the likelihood of discovering more sensitive information, potentially different from what was discovered by previous authors (Vasilescu et al. 2015b). This selection resulted in a final corpus of 15,672 comments. However, since these comments in the GHTorrent dataset were truncated, we updated all of them through the GitHub API. This process in represented in Fig. 10. This corpus, developed with the aid of the Privacy Dictionary, was manually labeled as described in the following Section 4.4 to classify sensitive comments and the type of information disclosed.

4.4 Manual Labeling Process and Protocol

After skimming the pull_requests comments corpus, as outlined in Section 4.3, we curated a set of 2,000 comments, representing nearly 10% of the extracted comments. To ensure a sufficient level of informativeness, we specifically chose comments with a minimum of 2,000 characters. The aim was to explore the nature of information that users potentially disclose in their pull_request comments. The selection of longer comments, coupled with

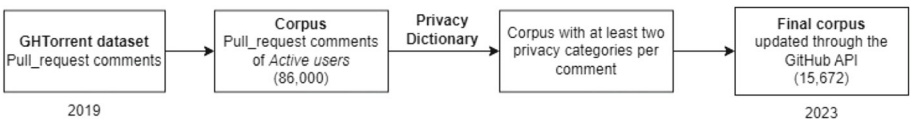


Fig. 10 Process for constructing the corpus

their prior labeling using the Privacy Dictionary, aimed to enhance the likelihood of filtering out irrelevant or purely technical comments. These comments were subjected to manual labeling by all the authors of this paper.

The annotation team consisted of four members, ensuring a gender balance. All annotators had STEM backgrounds and held various academic positions, ranging from PhD student to full professor. Three of the annotators were from the same country, while the fourth was from a different one. Each annotator was given a file of roughly 1,000 comments; every file was assigned to at least two different annotators, following existing guidelines about how to conduct a user study (Di Rocco et al. 2021; Robillard et al. 2010). A value of 1 was assigned to comments that revealed personal information about the user, 0 otherwise, irrespective of our perception of sensitivity. Sensitivity is a subjective concept influenced by social and cultural factors; thus, what one deems sensitive may differ from person to person. Indeed, for the manual labeling process, we refer to the definition provided by Bioglio and Pensa (2022) about Privacy-sensitive content:

> A generic user-generated content is privacy-sensitive if it discloses, explicitly or implicitly, any kind of personal information about its author or other identifiable persons.

For comments deemed as disclosing information, annotators had to select a label from the **Possible category** column corresponding to the type of information disclosed. The proposed labels were **Personal name**, **Workplace**, **Email**, **Location**, and **Gender**. The labels regarding personal name, job's information, email and location were derived from a preliminary corpus analysis conducted by one of the authors. Furthermore, these labels aligned with the privacy settings available on the GitHub profile, where users have the option to conceal specific private information. The inclusion of the gender label was prompted by findings from various studies on GitHub that highlight instances of gender discrimination or non-inclusive behaviors (Imtiaz et al. 2019; Garcia et al. 2023). Annotators were allowed to choose multiple labels for each comment and could use the **Other** column to indicate information not covered by provided labels. Participants were given an annotation guide to enhance consistency throughout the process.

After completion, the four annotators met together via Teams in two different days to discuss the comments marked as privacy sensitive and to reach an agreement in case a comment was marked only once. Each comment expressing disagreement underwent thorough discussion between the two assigned annotators, while the remaining two played a moderating role, facilitating the conversation and aiding in reaching a consensus. During these discussions, one annotator was prompted to explain why a comment was considered sensitive or not sensitive to the other annotator of the same comment. Agreement would result in either flagging the comment as sensitive or deleting it. Moreover, we discovered during these meetings that several sensitive comments disclosed information not covered by the existing labels. This information was added by the annotators in the column **Other**. We unanimously decided to introduce additional labels, such as **Moral values**, **Community etiquette**, **Personal info**, **Language**, and **Relation with a user**, to account for instances where comments disclosed such information. Ultimately, 147 comments were unanimously identified as disclosing private user information. Table 9 provides an excerpt from the fully annotated corpus, displaying comments with specific labels and brief descriptions of the label meanings.

On the labeled corpus, we conducted an analysis to calculate the level of agreement. Each comment was annotated by two annotators, resulting in two raters per comment. Across the entire dataset, we computed Cohen's kappa coefficient, a metric utilized to assess inter-rater reliability (Cohen 1960). The Kappa score of 0.49 was computed for the binary sensitivity column before any discussion occurred, indicating a moderate level of agreement among the

Table 9 Examples of labels assigned to a comment and description of the label's meaning

Comments	Label	Description of label
1) I tried looking for you and it showed 2 different people with that same email address and when I went to start a conversation with them it said neither had Google Hangouts. :\My email is [EMAIL] - always on Hangouts	Email	Comments containing user's email.
2) On doit pouvoir le faire de façon un peu plus générique... en demandant si la propriété existe. Quelque chose du style: if(getProps().hasProperty(kFnOfxParamPropGroupIsTab) getProps().propSetInt(kFnOfxParamPropGroupIsTab, 1);	Language	Comments in which a user's writes in a language different from English.
3) Is it just me who expected to see the full text from which the "During the ... quite a bit [...]" quote was extracted upon clicking the headline link for this item? Having a direct link to the release download page is wonderful, but can we also have a link somewhere to the "announcement" the text was taken from? Or is this not an excerpt from anything but an original text?	Moral values	Comments in which there is a moral consideration made by the user that discloses user's ethics or moral values.
4) lib @[$USER_1$] @[$USER_2$] @[$USER_3$] @[$USER_4$] My activity on UMS will be sporadic over the next few weeks, as I have family from overseas visiting. I'll still try to do what I can. Do you think we should release 3.4.0 now?	Personal information	Comments containing any type of personal information not represented by the other labels, for example: information about user's family, trips, studies, habits and so on.
5) This is a good start. Our telephone conversation makes more sense now. I don't think we'll be able to reuse _exactly_ same container ('TeamProjectCardContainer') in the meeting context - that TeamMembers subscription on the container doesn't exactly make sense. We'll probably have to write another container component. This branch is a good illustration of what we need to do. I didn't notice the 'isProject' prop on 'OutcomeCard' for example.	Relation with a user	Comments containing information that reveal a direct relationship with another GitHub user.
6) @[$USER_5$] This is still a very young project. In the latest release we changed the command line syntax as well as I added a 'v' to the version number when tagging. I would have expected more of an outcry from the former rather than passive aggressive comments about the latter. I'm sorry we don't live up to your high standards, but I fail to see how your input here brings the project forward in	Community etiquette	Comments containing remarks on how to behave on the GitHub platform.

Table 9 continued

Comments	Label	Description of label
any constructive manner. Many other projects don't tag at all, fail to upload tarballs, or if they do said tarballs don't build, and most other projects don't even provide a changelog. Respectfully, comment on things that actually matter.		
7) Released. On Fri, Jan 17, 2014 at 12:40 PM, [PERSONAL NAME OF A USER] notifications@github.comwrote: Any chance you would release a new version? I just spent two hours tracing down this same bug. Thanks. Reply to this email directly or view it on GitHub [LINK TO A GITHUB PAGE] [PERSONAL NAME OF A USER] [COMPANY WEBSITE]	Workplace	Comments revealing information about user's job or workplace.
8) Basically you can attach release notes to your tags: [LINK TO A GITHUB PAGE] [PERSONAL NAME OF A USER] Den 08/02/2015 kl. 22.06 skrev [PERSONAL NAME OF A USER] notifications@github.com: I haven't checked out that feature. What's the scoop? On Feb 8, 2015, at 11:54 AM, [PERSONAL NAME OF A USER] notifications@github.com wrote: I thought we were switching to the releases feature on GitHub? @[$USER_6$] [LINK TO A GITHUB PAGE] @[$USER_7$] [LINK TO A GITHUB PAGE] Reply to this email directly or view it on GitHub [LINK TO A GITHUB PAGE]. Reply to this email directly or view it on GitHub.	Personal name	Comments showing personal name of the user (not their account name).
9) Yeah, we're good. I'll see if I can get posterior prediction with new levels working for real. On Sun, Dec 27, 2015 at 7:51 AM, [PERSONAL NAME OF A USER] notifications@github.com wrote: Hope you and your family avoided the worst of the tornado. On Sun, Dec 27, 2015 at 12:48 AM, bgoodri notifications@github.com wrote: @[$USER_8$] [LINK TO A GITHUB PAGE I pushed a new_thing() function, which is not complete but works like lme4's predict.merMod(). It sort of works but needs to be properly integrated into posterior_predict. I would finish it off, but I am about to be hit by a tornado. Reply to this email directly or view it on GitHub [LINK TO A GITHUB PAGE].	Location	Comments reporting any information that can be linked with the user's location.

raters (Warrens 2015). This is not surprising considering the novelty of the analyzed data and the nuances present in the dataset. It could also be attributed to the observation that, out of 2,000 comments, 1846 received identical binary labels from both evaluators.

5 Sensitivity Detection in Textual Comments

Pre-trained generic language models (Devlin et al. 2018; Inan et al. 2023; Howard and Ruder 2018) have achieved great results on different NLP tasks. To illustrate the potential of these models in enhancing user privacy awareness, our study concentrates on demonstrating their ability to detect possible privacy leaks within textual comments. The primary aim is not to develop high-performance tools or methodologies but rather to explore the feasibility of autonomously detecting self-disclosure across various contexts.

To achieve this, we fine-tuned the Llama2 model (Touvron et al. 2023) (Section 5.1), using the labeled corpus of comments obtained through the process explained in Sections 4.3 and 4.4. This fine-tuning process aimed to enhance the model's performance in identifying privacy data leakage, leveraging the targeted information in the curated dataset.

Additionally, we investigated the performance of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2018) for sensitivity detection (Section 5.2). BERT, a widely used pretrained model, was fine-tuned for binary classification of self-disclosure in comments. We utilized the Hugging Face Transformers library and trained BERT on our labeled dataset curated in Section 4.4 to adapt it to the sensitivity detection task. The task consists of a binary classification of sensitive or insensitive comments.

5.1 Large Language Models for Sensitivity Detection

Research in the domain of content sensitivity detection and privacy leakage in text spans various disciplines, including machine learning (ML), natural language processing (NLP), philosophy, psychology, and the social sciences. The overarching goal is to enhance privacy awareness and conduct risk assessments, primarily within specific platforms, with the ultimate aim of developing technology for empowering users in protecting their privacy. Despite the considerable success of these approaches, they generally do not conform to a one-size-fits-all model (Nguyen et al. 2023; Tang et al. 2023). Performance disparities exist across datasets, with models excelling in specific contexts while underperforming in others, as demonstrated in experiments by Peiretti and Pensa (2023). Language, dataset balance, and text length are among the factors influencing model effectiveness. Moreover, achieving satisfactory performance necessitates a harmonious combination of feature extraction and classifier design. Researchers are required to explore numerous combinations to optimize synchronization (Nguyen et al. 2023; Tang et al. 2023). This comprehensive approach is crucial for content sensitivity detection in text. Recent advancements in LLMs have significantly enhanced the performance of diverse natural language processing (NLP) tasks (Min et al. 2023), opening up new possibilities for automating functions traditionally executed by humans. These models consist of large neural networks pretrained on vast corpora of text data in multiple languages, offering potential solutions to the challenges associated with conventional text classification methods. This is the reason why we chose to use LLMs to for this detection sensitivity task. In order to demonstrate the potential of Llama, we have asked the model to provide only Yes or No as an output. In particular, Listing 1 shows this instruction.

Listing 1 Zero-shot Llama model.

```
def query_llama(comment: str) -> str:
    return f"""### Instruction: We need to classify the text as
    privacy sensitive or not, and please use Yes or No.

### Input:
{comment.strip()}""".strip()
```

Several successful research endeavors have utilized fine-tuning of large language models (LLMs) (Lu et al. 2023; Behnia et al. 2022; Yao et al. 2024). In order to enhance the accuracy of Llama's predictions, we have chosen to utilize the Parameter-Efficient Fine-Tuning (PEFT) method. It allows for the effective customization of pre-trained language models (PLMs) for different downstream applications without the need to fine-tune all of the model's parameters. Optimizing extensive pre-trained language models (PLMs) is frequently too expensive. PEFT approaches specifically focus on fine-tuning a limited number of additional model parameters, resulting in a significant reduction in both computational and storage expenses. The fine-tuning process involves providing a cue to the model and directing it to provide an appropriate binary classification, specifically distinguishing between privacy-sensitive and privacy-non-sensitive. The target provided corresponded to the projected classification. This was done to enable the model to provide a direct response using binary classification. The prompt employed during the process of fine-tuning is illustrated in Listing 2. It consists of three main parts: *(i) Instruction*, where we are asking for a binary classification; *(ii) Input* is the comment to be classified; and *(iii) Response* is the expected answer. The results of this experiment are illustrated in Section 6.

Listing 2 Fine-tune the Llama model with the labeled corpus.

```
def generate_training_prompt(comment: str, bin_class: str) -> str:
    return f"""### Instruction: We need to classify the text as
    privacy sensitive or not, and please use Yes or No.

### Input:
{comment.strip()}

### Response:
{bin_class}
""".strip()
```

5.2 BERT for Sensitivity Detection

In our study, we explore the capability of BERT to classify user privacy disclosures effectively. Its architecture is based on a transformer encoder and the basic BERT architecture consists of different attention-based layers. The tokens are represented as input vectors which include the tokens they-self, their positions, and their context sentence. For each token, the attention-based layers produce a representation. Each token representation is based on the representations of all tokens. The output of one attention-based layer is provided as input for the next one. Finally, the last attention layer provides the model output. The BERT model is unsupervisedly trained on large amounts of text and may later be applied to potentially any task. This allows us to train the `bert-base-uncased` model[7] using the labeled set of comments defined in Section 4.4. We have chosen this model because previous authors have successfully employed BERT in the context of empowering users (Adhikari et al. 2022; Khalajzadeh et al. 2022a; Wang et al. 2022). The applications span from automated labeling of GitHub issues to privacy policy classification.

[7] https://huggingface.co/google-bert/bert-base-uncased

5.3 Metrics and Methodology

Our study evaluates the effectiveness of pre-trained models in identifying private information within text by conducting three distinct configurations.

The first configuration, zero-shot Llama ($\texttt{Llama}_{zs}$), employs the LLaMA model utilizing predefined queries as outlined in Listing 1. This approach tests the baseline ability of LLaMA to recognize privacy-related information without any model customization. In the second configuration, fine-tuning Llama ($\texttt{Llama}_{ft}$), we enhance the LLaMA model's capability by fine-tuning it on 80% of our curated dataset as outlined in Listing 2. The remaining 20% of the data serves as a test set to evaluate the model's prediction accuracy. The third configuration, fine-tuning BERT ($\texttt{BERT}_{ft}$), involves a BERT model that we trained specifically for the task. Similar to the $\texttt{Llama}_{ft}$ configuration, we fine-tuned the model on 80% of the manually curated comments, reserving the remaining 20% for performance evaluation.

To asses the performance of the model, we used state-of-the art metrics: accuracy, precision, recall, and false positive rate (FPR). In what follows, *TP* is the number of true positive predictions, *TN* is the number of true negative predictions, *FP* is the number of false positive predictions, and *FN* is the number of false negative predictions. The metrics are calculated as follows (Hossin ans Sulaiman 2015):
Accuracy refers to the ratio of right predictions, including both true positives (TP) and true negatives (TN), to the total number of cases analyzed:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP+TN+FP+FN}} \tag{1}$$

Precision measures the fraction of the number of correctly classified comments as yes (TP) to the total number of classified comments as yes (TP + FP):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Recall measures the proportion of actual positive cases that are correctly identified by the model. It is defined as the ratio of true positives (TP) to the sum of true positives and false negatives (FN):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

F1 represents the harmonic mean between recall and precision values, and it is particularly high when true negatives (TN) are high:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Section 6.4 shows the results of the different experiments and compare the performances of each configurations.

6 Empirical Results

In this section, we report and analyze the experimental results by answering the four research questions introduced in Section 1.

6.1 RQ$_1$

Are the privacy settings provided on GitHub used/adopted by the users? In other terms, do we observe different combinations of these settings, or is there a dominant configuration?

Users' dataset cluster analysis The cluster analysis conducted on the *Users'* dataset led to three unbalanced clusters, shown in Fig. 11a. Figure 11b, c and d show the number of users that set (1) or hide (0) the corresponding setting being `City`, `Company`, and `Longitude`. From these plots, we can observe that the clusters are distinguished according to each variable considered. For example, users in the first cluster, named "Concerned", exhibit a reluctance to share any information on GitHub. Conversely, those in the second cluster, denoted as "Average Concerned" are willing to share information only regarding their location (`Longitude`). On the other hand, users in the third cluster, namely "Unconcerned", have a tendency to share comprehensive information through their privacy settings. Indeed, the majority of the users in this cluster have set the option to disclose both their `City`, their `Company`, and their

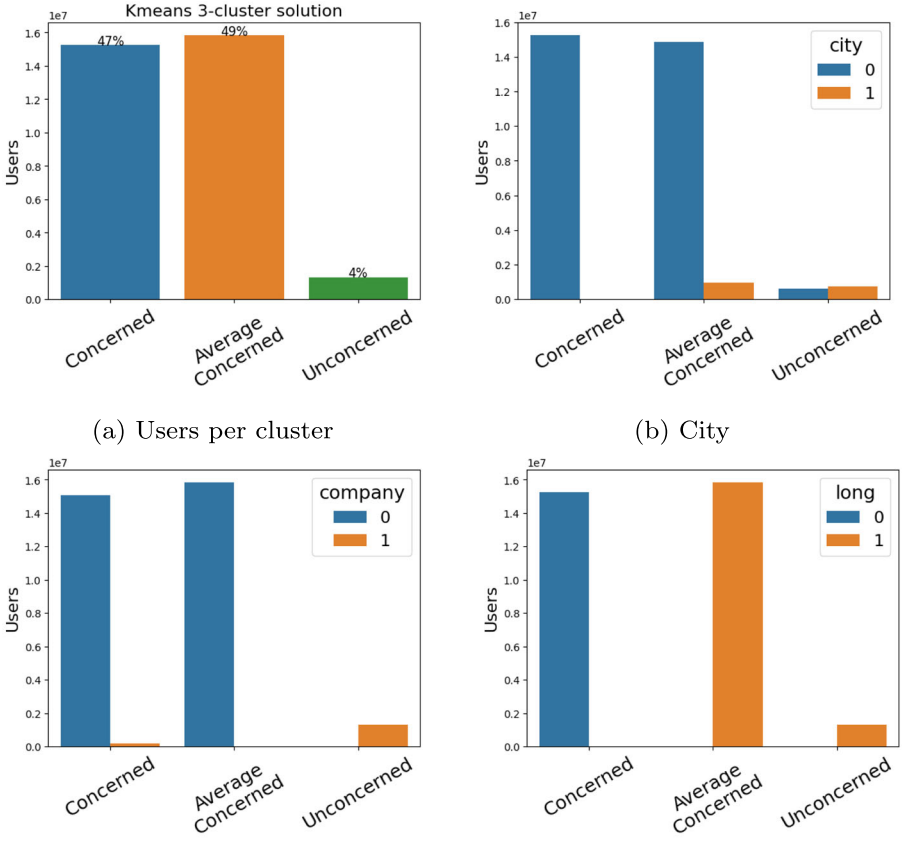


Fig. 11 Cluster analysis on the *Users'* dataset with K-means. Number of users per cluster (Fig. 11a) and distribution of variables per cluster (Fig. (11b, c and d)

`Longitude` (location). This solution suggests that the privacy desiderata of GitHub users can vary even on a small set of privacy options and it demonstrates that GitHub users actively utilize privacy settings. It is interesting to observe the population distribution depicted in Fig. 11a, as the cluster "Unconcerned" stands out as the less populated cluster (4% of the entire population). This observation potentially implies that only a limited percentage of GitHub users are willing to disclose comprehensive information in their profiles. This finding further motivates our study.

Active users analysis with K-means On the *Active users* dataset, we performed a K-means cluster analysis with K=4 chosen with the Elbow method. Figure 12 shows an overview of privacy settings choices made by the active users. The privacy profiles are rather balanced, as illustrated in Fig. 12a by the cardinality of each cluster. Figure 12b, c, d, e, and f depict users' privacy settings choices per cluster. We used this analysis to qualitatively define each profile as follows: "Concerned" about hiding their information from the GitHub profile, "Little concerned", "Unconcerned" and "Average concerned". For instance, Fig. 12b reveals that in the first cluster, namely "Concerned", a significant majority of users opted to conceal their `City` information on their profiles, similarly to users in cluster "Average concerned". On the contrary, users from clusters "Little concerned" and "Unconcerned" exhibit a significant number of users willing to share information about their City. Analogous considerations can be applied to the other variables: `Company`, `Email`, `Events` and `Twitter`.

These results are significant as they illustrate that users' privacy concerns, as expressed through the privacy settings of GitHub, vary. Thus, users exhibit distinct privacy profiles. This is visible in Fig. 12a, where the profiles are rather balanced in terms of the number of users, showing that the choices of privacy settings do not converge towards one main combination of settings. These privacy profiles helped us categorize active users according to their privacy desiderata. We exploited this categorization to address RQ_3.

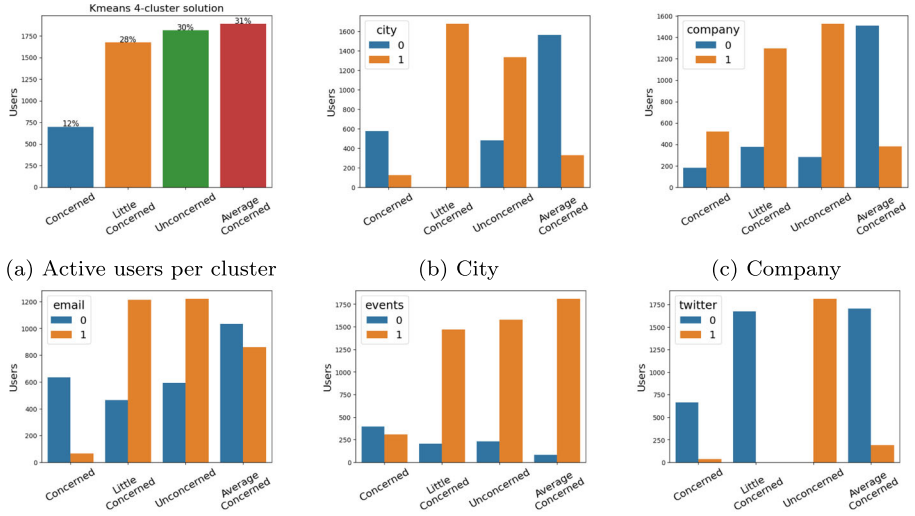


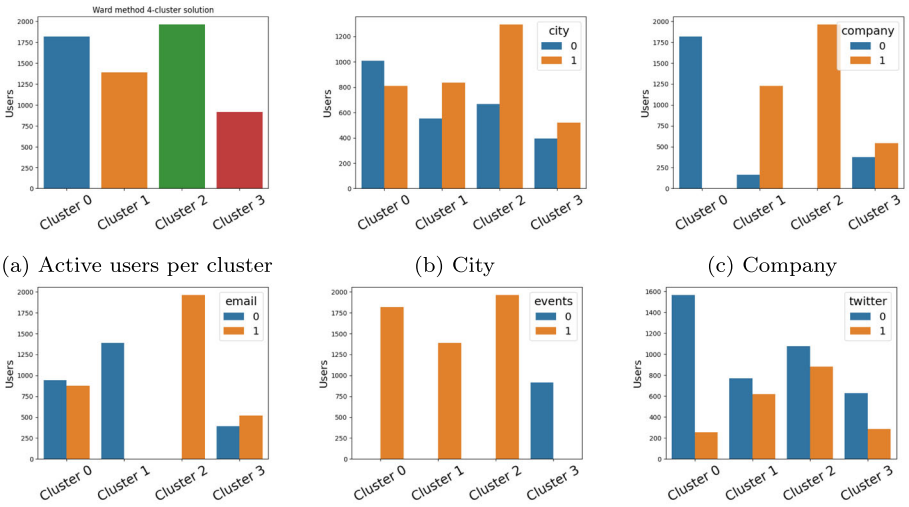
Fig. 12 Cluster analysis on the *Active users* dataset with K-means. Number of active users per cluster (Fig. 12a) and distribution of variables per cluster (Fig. 12b, c, d, e and f)

Active users analysis with hierarchical clustering As many authors suggest (Brandäo et al. 2022; Sanchez et al. 2020), another way to generate users' profiles is through a hierarchical clustering algorithm. We used this method to further analyze the *Active users* dataset. We applied this technique using Ward's method on the *Active users* dataset, with the number of clusters equal to 4. As previously explained, this number was chosen by analyzing the dendrogram (Fig. 9). We report the bar charts regarding the distribution of variables per cluster and the cardinality of each cluster (see Fig. 13). The clusters are less balanced than the one obtained with K-means (Fig. 13a). By observing the distribution of variables per cluster, the situation is unclear compared to the clusters with K-means. Indeed, the variable `City` seems irrelevant in discriminating between the profiles (Fig. 13b).

Answer to RQ$_1$. GitHub users manifest diverse privacy preferences, as reflected in their selection of privacy settings, both on a broad scale -exemplified by the analysis conducted on the entire *Users* dataset- and on a more granular level, as seen in the examination of the *Active user* dataset. The latter is interesting, given that these users share a high level of activity on the platform, yet their privacy preferences can vary considerably. According to our analysis of the *Users* dataset, it emerges that only a small percentage of users are willing to disclose all their information in their GitHub profiles. Overall, privacy settings are a tool used by users to safeguard their privacy and should accurately reflect their privacy preferences.

6.2 RQ$_2$

What types of private information are disclosed on GitHub by users?



(a) Active users per cluster (b) City (c) Company

(d) Email (e) Events (f) Twitter

Fig. 13 Cluster analysis on the *Active users* dataset through the hierarchical clustering methods. Number of active users per cluster (Fig. 13a) and distribution of variables per cluster (Fig. 13b, c, d, e and f)

In order to address RQ_2, we began with a dataset of 2,000 texts of *Active users*, selected using the Privacy Dictionary, as described in Section 4.3. From this corpus, we manually labeled 147 comments as privacy sensitive.

This corpus provides examples of different types of private information that is disclosed by the GitHub users. Sometimes this information is more explicit, as visible in comment 1) from Table 9. In this case, the user reveals his/her personal email. Sometimes, a piece of private information is more implicit, as in comment 8), where the user speaks of a tornado in the area of his/her interlocutor. This reveals a close relationship between the users and the information of the tornado can lead to the location of the interlocutor. Similarly, comment 5) reveals a close relationship between the two users, having a conversation on the phone. Figure 14 shows the percentage distribution of each label in the corpus. As it is evident, "Personal name" represents a significant portion of the pie chart. However, it is noteworthy that nearly 40% of the sensitive comments contain a wide range of sensitive information, from "Moral values" to work-related details.

This finding reaffirms the observations made by previous researchers, that while GitHub primarily serves as a platform for sharing technical knowledge, an examination of users' textual comments expose instances of disclosing private information. Such information may originate from the commenter, either pertaining to themselves or involving another user.

Answer to RQ_2. Even if GitHub is considered a platform used for technical purposes only, different types of private information are consciously or unconsciously disclosed by the users. The categories of information revealed in pull_request comments exhibit diversity. To date, our investigations have uncovered instances of **Personal Name**, **Workplace**, **Email**, **Location**, **Moral Values**, **Community Etiquette**, **Personal Info**, **Language**, and **Relation with a User**.

6.3 RQ_3

After users choose their privacy settings, do they adhere to what they have declared? In other words, can we observe a discrepancy between their stated privacy preferences and their

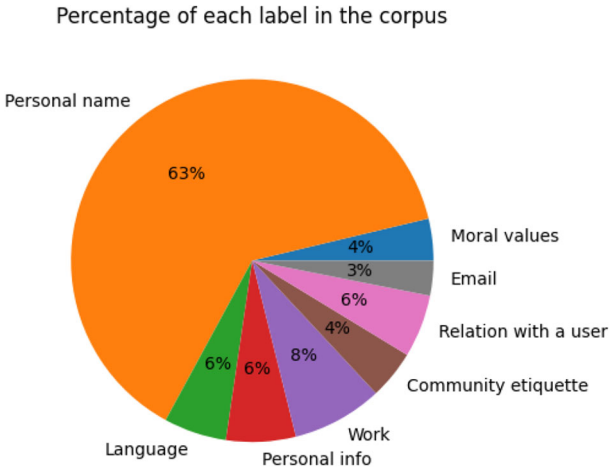


Fig. 14 Percentage of each label in the corpus

actual behavior, such as their textual activity?

During the manual-labeling process described in Section 4.4, we selected comments that were particularly meaningful from a privacy perspective to analyze the profile of the author. Table 10 presents examples of sensitive comments and the profiles of their authors, as identified in Section 6.1.

Interestingly, many of these comments fall into the profiles of "Average Concerned" and "Concerned" users. This suggests that users who are presumed to be concerned about their privacy do not necessarily demonstrate this concern through their behaviors. Figure 15 shows the distribution of each label per privacy profile, i.e., how many comments disclosing that information were found in each cluster. The bar plot in Fig. 15a represents the distribution of sensitive comments across different privacy profiles. Each privacy profile is indicated on the x-axis, categorized into the four groups: "Concerned", "Little Concerned", "Unconcerned", and "Average Concerned". The bar plot in Fig. 15b presents the same data with the y-axis logarithmically scaled, allowing for a more compact and interpretable visualization. Consistent with previous findings, the label "Personal name" is the most commonly shared across all privacy profiles. As shown in the plot, the profile of "Concerned" users displays six out of nine labels, which is expected as these users are likely more attentive to their privacy. However, evidence of self-disclosure is still found among users in this cluster. Conversely, the "Average Concerned" user profile contains all the different labels, suggesting a discrepancy between their stated privacy preferences and their actual behaviors.

This phenomenon can be interpreted in various ways. One possible explanation is that GitHub privacy settings may not comprehensively capture users' privacy preferences. Alternatively, developers might believe that disclosing certain information could be advantageous in specific situations, potentially overlooking the fact that this information is openly available. Another consideration is the so-called "privacy paradox" (Acquisti and Grossklags 2005; Gerber et al. 2018). For example, users in the profile "Average Concerned" express a desire to keep information private through their privacy choices but exhibit a disclosing behavior in their writing/commenting activities. However, the validity of the privacy paradox is highly debated in the literature, and its existence is questioned (Solove 2021). Lastly, this finding can be also explained with the concept of "privacy fatigue", which refers to the increasing difficulty individuals face in managing their online personal data, leading to weariness about having to constantly consider online privacy (Choi et al. 2018).

Answer to RQ_3. The privacy preferences selected by the users on GitHub might not be entirely representative of their privacy desiderata. Indeed, we found a discrepancy between what they declared as privacy settings, and how they behaved. This finding can be attributed to various factors, spanning from a perceived advantage in sharing specific information, to the so-called "privacy fatigue", to a potential lack of awareness, often referred to as the "privacy paradox".

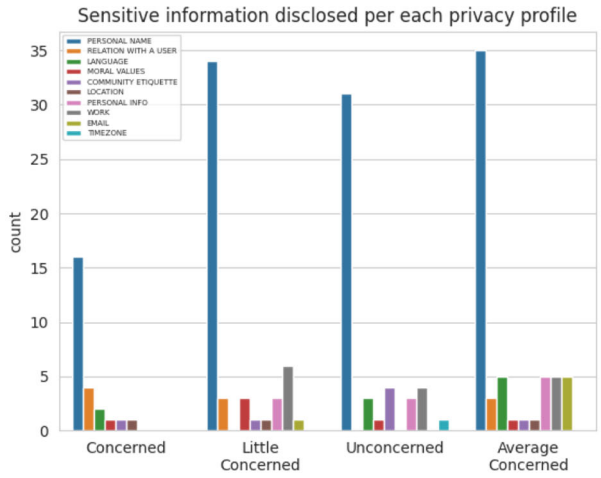
6.4 RQ_4

To which extent is it possible to automate the detection of sensitive comments with the use of BERT or Llama2?

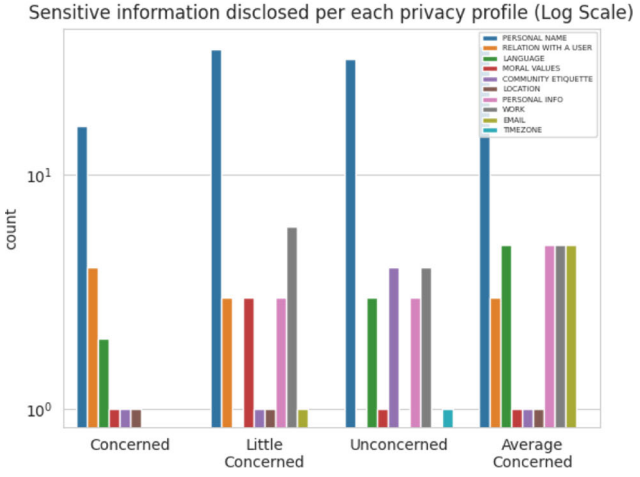
This section compares the configurations described in Section 5.3, each fine-tuned or used directly to classify privacy disclosures in textual comments. Our analysis focuses on

Table 10 Examples of comments in each profile

Comments	Privacy Profile
My two cents. This sure would be a very different conversation if it was 'WebM for Niggers' or 'WebM for Kykes'. Those words are more accepted as being derogatory in society. I think a lot of people use 'retard' as a colloquial, playful word, but really, I think it isn't acceptable; a community, however minority, is [voicing it's opinion]([LINK TO A WEBSITE]]) more openly about this. I think they are right. We must, I believe, be wary and think critically about any form of censorship. In this case, I believe the censorship is legitimate because the change does not affect any functionality of the repository or suppress opinions (blowing up the forks was quite silly though). Each case of censorship should be exposed to this amount exposure and community thought. I think we are doing well in that regard. Github are enforcing their TOS and I believe this change is for the positive. People implying that Github are restricting their freedom, I think, will need to try and empathise more with the community of people affected by this word.	CONCERNED
@[USER$_7$] Hey Gil, I am sorry if my comment sounded harsh to you. I know you are working your ass off for the community and I really appreciate your work here. Applying changes to all our packages is a boring and unthankful job, I can tell that from own experience. I understand that you don't have time to create pull requests in that situation and I am sorry that I did not take this into account when asking for a PR. I got frustrated lately with plone.restAPI development a bit because other devs with good intentions do things that lead to additional work for me as a maintainer. plone.restAPI uses semantic versioning and my goal is to do very frequent releases after every single meaningful pull request. This is only possible if I can merge PRs at any time and then do a release right away. This wasn't possible because of your commit to master. With a PR I could have easily postponed the merge of this important and valuable contribution to after the next release. The simple reason why I did not complain about Maurits PR is that this one did not require me to manually go through all open PRs and having to amend them.	LITTLE CONCERNED
A lot of thought and effort has gone into researching the subject of "deep overriding" by Russell O'Connor in the 'haskellPackagesFixpoint' branch and by myself in the 'haskell-ng' branch. You are aware of these activities, but apparently chose to ignore them and instead committed your own "deep overriding" solution to 'master' without any prior consultation. There is nothing inherently "wrong" with doing that; you are entitled to commit whatever changes you feel are best. Personally, I perceive this kind of unilateral decision making as disrespectful towards other stakeholders, though. Also, I feel that your solution is weird from a technical point of view, and having it reviewed by others before you committed certainly wouldn't have hurt the quality/readability/re-usability of your patch.	UNCONCERNED
Hi [PERSONAL NAME OF A USER], Of cause I follow your development. I have a script running which checks genode.org for changes. The release notes are like early Christmas presents :) I would like to go to FOSDEM next year, but I'm not sure if Daniel will be able to come with. Are you attending? It would be amazing to meet up with you guys again. Cheers /u	AVERAGE CONCERNED
To be honest, I just missed it when I was reviewing the release notes @[USER$_8$] had created in the first place. I don't mean to be political hence the removal since whether we like it or not that phrase brings certain things to mind.	



(a) Distribution of label in each privacy profile.



(b) Distribution of label in each privacy profile with the y-axis logarithmically scaled.

Fig. 15 Analysis of the disclosure of each label per privacy profile

several key performance metrics and explores the models' practical effectiveness in real-world privacy identification tasks.

In Table 11, we present the results of configurations in which comments were classified as either privacy-sensitive (PS) or non-sensitive (PNS). In particular, the table compares the three configurations described in Section 5, i.e., $\texttt{Llama}_{zs}$, $\texttt{Llama}_{ft}$, $\texttt{BERT}_{ft}$, across four metrics described in Section 5.3. Evidently, by the $\texttt{Llama}_{ft}$ and $\texttt{BERT}_{ft}$ configurations, the prediction performance is better than that of $\texttt{Llama}_{zs}$. The accuracy metrics stand at 0.94, 0.94 and 0.417, respectively, signifying that the fine-tuned configurations outperform the zero-shot in predicting the correct class. Across all the configurations, it is noticeable that the values for predicting non-sensitive instances consistently outweigh those for sensitive ones.

Table 11 Prediction performances for binary classification

	$\texttt{Llama}_{zs}$			$\texttt{Llama}_{ft}$			$\texttt{BERT}_{ft}$			
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1	Support
PS[1]	0.69	0.35	0.46	0.59	0.38	0.46	0.52	0.42	0.47	27
PNS[2]	0.79	0.42	0.55	0.96	0.98	0.97	0.96	0.97	0.97	373
Acc.	0.42			0.94			0.94			400

[1] PS stands for the privacy sensitive class
[2] PNS stands for the privacy not-sensitive class

This imbalance can be attributed to the fact that the dataset used is heavily skewed towards non-sensitive texts. Further improvement can be achieved with a more balanced dataset.

Even though we have asked the model to provide only "Yes" or "No" as acceptable answers (see Listing 1), we observed that the output of the $\texttt{Llama}_{zs}$ was frequently uninterpretable. Examples such as "I'm not sure what you mean by 'privacy-sensitive,' I'm not sure what you mean by 'not'" and "This is a simple yes or no" were generated by the model. On the contrary, the $\texttt{Llama}_{ft}$ always generated interpretable output, with a clear answer to the prompt given which consisted in either "Yes" or "No". Further studies should delve into prompt-engineering techniques to assess whether the performance in predicting sensitive text can be significantly enhanced. It is worth noting that by $\texttt{Llama}_{ft}$ and $\texttt{BERT}_{ft}$ their prediction performance is somehow comparable, two indicating that both models can be exploited for building a tool of sensitivity detection. The curated dataset together with the python scripts to preprocess and fine-tuning pre-trained models are available on the supporting GitHub repositories.[8]

Answer to RQ$_4$. Pre-trained models can be employed to identify sensitive information in textual data. The availability of a curated corpus of user comments enables the fine-tuning of pre-trained models. Both $\texttt{Llama}_{ft}$ and $\texttt{BERT}_{ft}$ configurations exhibit superior performance compared to the $\texttt{Llama}_{zs}$. Nevertheless, to ascertain whether the low performance on sensitive data is attributed to the skewed dataset or the model itself, it is recommended that more examples of sensitive texts be introduced for evaluation.

7 Discussion

Our empirical analysis addressed the study of the privacy dynamics on GitHub, with particular attention to users' privacy settings and behaviours on the activities related to pull_requests comments. Users privacy preferences were deduced from the privacy settings they chose in their profiles, while the analysis of their behaviours was conducted on their textual activity (pull_requests comments). Primarily, we observed that users from both *Users* and *Active users* dataset exhibit significantly distinct privacy preferences. This finding indicates that users actually adopt privacy settings and that they express different privacy concerns (RQ1). Consequently, there is a clear indication for a more thorough investigation into the dynamics of privacy on this platform.

Additionally, despite GitHub being primarily used for sharing technical knowledge, there are instances of unintentional or deliberate leakage of users' private information in their textual activity (RQ2). The disclosed information ranged from real names to personal values,

[8] https://github.com/MDEGroup/EMSE-CHASE-Privacy

surpassing what could be concealed by the privacy settings provided on GitHub. This is in contrast with the GitHub privacy statement, which asserts that it is sufficient to "adjust your setting for your email address to be private in your user profile" (see Fig. 1). Indeed, along with previous studies (Vasilescu et al. 2015a; Terrell et al. 2017; Meli et al. 2019; Niu et al. 2023), we realized that this statement does not hold true. This suggests that privacy settings alone do not guarantee users' privacy and that a more sophisticated tool for privacy protection and awareness is needed.

The analysis of user behaviors (pull_request comments) has enabled us to identify diverse types of sensitive information disclosed on GitHub and compare them with the privacy preferences expressed by the users (privacy profile). We observed that users assigned to a privacy-concerned profile were authors of privacy-sensitive comments (RQ3). Our findings indicate that although users do engage with privacy settings using various configurations, their behaviors may inadvertently expose certain private information. This can be attributed to users' lack of awareness, convenience, or privacy fatigue. In any case, more sophisticated privacy settings on GitHub would allow users to more accurately reflect their preferences. Due to the limitations of privacy settings, we explored using Llama2 and BERT to detect sensitive comments on GitHub (RQ4). This preliminary study suggests that BERT outperforms Llama2 in this task. Further investigation with more prompt engineering should be conducted to confirm this result. The implementation of finer-grained privacy settings could serve as a foundation for developing a privacy awareness tool on GitHub, which could notify users when their behavior, identified with the help of models like Llama2 or BERT, deviates from what is specified in their profile. In this context, a privacy awareness tool could be useful for alerting users when such deviations occur or for suggesting less sensitive rephrasings of text.

7.1 Threats to Validity

– **External validity.** By using the GHTorrent dataset, we were able to get a large sample of privacy settings on GitHub, as well as the number of actions performed by each user. This allowed us to establish a definition of *Active users* on the platform. It is worth noting that the GHTorrent version we used was from 2019, which means our selection of *Active users* was based on data available up to that time. So, no new users were added to the original GHTorrent data and the analysis was done on users who were part of the 2019 dump. To tackle this potential limitation, we updated users' privacy preferences and their comments using the GitHub API. Moreover, our results mainly concern the population of *Active users* and might not apply to other GitHub users. Future studies could address this limitation by directly collecting data from the platform and verifying whether the results still hold.
We have excluded comments without privacy-related labels from the fine-tuning and evaluation datasets. While this does not seem to compromise the model's ability to distinguish between sensitive and non-sensitive comments, the lack of these comments may limit the generalizability and robustness of the findings. A more diverse dataset could help address this concern.
The use of truncated comments from GHTorrent may have led to the omission of privacy-related terms in the truncated sections. While this does not affect the accuracy of the automated labeling, it may have resulted in an underestimation of the total number of privacy-sensitive comments.
The textual data comprised only pull_requests comments. We chose this as "pull request comments are likely to contain valuable insight into the relationships of developers inter-

acting with one another" (Sajadi et al. 2023), thus we expected to find more sensitive information in this type of data. Future work may consider also issues and commits.

– **Internal validity.** We adopted privacy settings chosen by the users as a declaration of their privacy desiderata. Even if this can be the case for some of them, for others their choice of privacy settings might be arbitrary or not necessarily aligned with their actual preferences. This discrepancy could stem from limitations in the options available on GitHub or from a lack of awareness regarding privacy implications. Moreover, some users might be unemployed and therefore not showing information related to their job, or not have a Twitter account. In our study, these cases are considered as hiding information.
– **Construct validity.** While there were four annotators in total, each comment was evaluated by only two individuals. People may perceive the same comment in different ways, and this can pose a threat to the corpus validity. To mitigate this bias, we organized the discussion phase, where any discrepancy was discussed and resolved by two involved participants. The agreement process took place in two different sessions, during which the raters could explain their choices.
  For a more nuanced understanding of user privacy, a method like the Experience Sampling Method could offer valuable insights (Zhang et al. 2020, 2021).

8 Conclusion and Future Work

With the aim of gaining a better understanding of users' privacy on GitHub, we conducted an empirical study on this platform regarding users' privacy preferences and behaviors. Our findings demonstrate that users actively engage with the platform's privacy settings, leading to the identification of four distinct privacy profiles that emerged from a cluster analysis on the privacy settings (RQ1). For what concerns privacy behavior, we found that users share different personal information on GitHub, including family details, location, and company-related information, among other things (RQ2). This indicates that on GitHub there is a wide range of private information that can be associated with a user, beyond technical matters and surpassing information that can be hidden using privacy settings.

The synthesis of these results revealed a discrepancy between the chosen privacy settings and the actual behaviors of users (RQ3). Despite the possible explanations for interpreting this phenomenon, it is evident that privacy settings alone are insufficient to ensure users' privacy. The last result underscores the necessity for more nuanced privacy settings on these platforms and the development of automated tools to assist users in consistently managing their actions on the platform. Given the limitations of privacy settings in protecting users' privacy, we explored adopting models like Llama2 and BERT to detect sensitive comments and pave the way for a privacy awareness tool (RQ4).

Indeed, our work enables the creation of a privacy awareness tool on GitHub that may advice users when they are entering sensitive information in their comments. In this direction, our corpus can be exploited for prediction on sensitive comments. At a more general level, this study lays the foundation for a methodology to observe similar privacy vulnerabilities also on other platforms.

In our future research, we endeavor to enhance our understanding of the necessity for automated privacy tools on GitHub. This will be achieved through the implementation of a user survey, akin to the one made by Vasilescu et al. (2015a). In this respect, it would be valuable to acquire updated data on privacy choices made by the same users in 2024. This could serve as a future work to explore the evolution of privacy attitudes and its potential

impact on the use of privacy settings. Furthermore, our objective is to augment the size of the labeled corpus by leveraging Llama2, as fine-tuned in this study, or BERT. We plan to utilize this corpus to create a multi-label tool capable of predicting specific sensitive information disclosed in comments. An expanded corpus should also include issue and commit comments to ensure a more diverse range of textual data is represented in the dataset. The various types of sensitive information disclosed on this platform can have varying consequences on an individual's life. In our future research, it is crucial to study the impact that each type of self-disclosed information may have on users' lives. This could also enhance the effectiveness of privacy awareness tools.

Ultimately, our overarching goal is to develop a tool that suggests sanitized comments upon identifying sensitive information thus empowering the users in managing their privacy concerns. The authors are actively progressing along this trajectory.

Acknowledgements This work has been partially supported by: MUR project 2020 "EMELIOT" grant n. 2020W3A5FY; MUR projects PRIN 2022 PNRR: "FRINGE: context-aware FaiRness engineerING in complex software systEms" grant n. P2022553SL, "TRex-SE: Trustworthy Recommenders for Software Engineers" grant n. 2022LKJWHC, and "HALO: etHical-aware AdjustabLe autOnomous systems" grant n. 2022JKA4SL. We also acknowledge the MUR Department of Excellence 2023 - 2027 program. We thank the anonymous reviewers for their useful comments and suggestions that helped us improve our manuscript.

Funding Open access funding provided by Università degli Studi dell'Aquila within the CRUI-CARE Agreement.

Data Availability Statements The experimental data and the simulation results that support the findings of this study are available in GitHub in the following address: https://github.com/MDEGroup/EMSE-CHASE-Privacy.

Declarations

Conflict of Interests The authors have no relevant financial or non-financial interests to declare.

References

Acar Y, Stransky C, Wermke D, Mazurek ML, Fahl S (2017) Security developer studies with GitHub users: Exploring a convenience sample. In: 13th Symposium on Usable Privacy and Security (SOUPS 2017), pp 81–95

Acquisti A Fong C (2020) An Experiment in Hiring Discrimination via Online Social Networks. Manag Sci 66(3):1005–1024. https://doi.org/10.1287/mnsc.2018.3269, https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2018.3269

Acquisti A, Grossklags J (2005) Privacy and rationality in individual decision making. IEEE Secur Priv 3(1):26–33. https://doi.org/10.1109/MSP.2005.22, https://ieeexplore.ieee.org/abstractdocument/1392696casa_token=rS6wHgIPjCQAAAAA:WAbt9Gq1MRK7TidTwlvgnrbn3MIftH6LzTnn8NiLPfW0pqPy8IuOQk8EEtZLD-sX30_agg

Adhikari A, Das S, Dewri R (2022) Privacy policy analysis with sentence classification. In: 2022 19th Annual International Conference on Privacy, Security Trust (PST), IEEE, pp 1–10

Alaei AR, Becken S, Stantic B (2019) Sentiment analysis in tourism: capitalizing on big data. J. Travel Res. 58(2):175–191

Autili M, Di Ruscio D, Inverardi P, Pelliccione P, Tivoli M (2019) A software exoskeleton to protect and support citizen's ethics and privacy in the digital world. IEEE Access 7:62011–62021. https://doi.org/10.1109/ACCESS.2019.2916203, https://doi.org/10.1109/ACCESS.2019.2916203

Bacchelli A, Beller M (2017) Double-blind review in software engineering venues: The community's perspective. In: 2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C), pp 385–396 . https://doi.org/10.1109/ICSE-C.2017.49

Barth S (2017) de Jong MD (2017) The privacy paradox – investigating discrepancies between expressed privacy concerns nd actual online behavior – a systematic literature review. Telematics Inf. 34(7):1038–1058. https://doi.org/10.1016/j.tele.2017.04.013, https://www.sciencedirect.com/science/article/pii/S0736585317302022

Becton JB, Walker HJ, Gilstrap JB, Schwager PH (2019) Social media snooping on job applicants: The effects of unprofessional social media information on recruiter perceptions. Pers Rev 48(5):1261–1280 . https://doi.org/10.1108/PR-09-2017-0278

Behnia R, Ebrahimi MR, Pacheco J, Padmanabhan B (2022) Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In: 2022 IEEE International Conference on Data Mining Workshops (ICDMW), pp 560–566. https://doi.org/10.1109/ICDMW58026.2022.00078

Bioglio L, Pensa RG (2022) Analysis and classification of privacy-sensitive content in social media posts. EPJ Data Sci 11(1):12

Blincoe K, Sheoran J, Goggins S, Petakovic E, Damian D (2016) Understanding the popular users: Following, affiliation influence and leadership on github. Inf Softw Technol 70:30–39

Blose T, Umar P, Squicciarini A, Rajtmajer S (2020) Privacy in Crisis: A study of self-disclosure during the Coronavirus pandemic. https://doi.org/10.48550/arXiv.2004.09717

Brandão A, Mendes R, Vilela JP (2022) Prediction of mobile app privacy preferences with user profiles via federated learning. In: Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy, pp 89–100

Casillo F, Deufemia V, Gravino C (2022) Detecting privacy requirements from user stories with nlp transfer learning models. Inf Softw Technol 146:106853

Chen Y, Zha M, Zhang N, Xu D, Zhao Q, Feng X, Yuan K, Suya F, Tian Y, Chen K et al (2019) Demystifying hidden privacy settings in mobile apps. In: 2019 IEEE Symposium on Security and Privacy (SP), IEEE, pp 570–586

Choi H, Park J, Jung Y (2018) The role of privacy fatigue in online privacy behavior. Comput Hum Behav 81:42–51

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46

D'Acunto D, Volo S, Filieri R (2021) "most americans like their privacy." exploring privacy concerns through us guests' reviews. Int J Contemp Hosp Manag 33(8):2773–2798

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Di Rocco J, Di Ruscio D, Di Sipio C, Nguyen PT, Rubei R (2021) Development of recommendation systems for software engineering: the CROSSMINER experience. Empir Softw Eng 26(4):69

Di Ruscio D, Inverardi P, Migliarini P, Nguyen PT (2024) Leveraging privacy profiles to empower users in the digital society. Autom Softw Eng 31(1):16

DiSalvo LM, Saenz GV, Wong WE, Li D (2022) Social Media Safety Practices and Flagging Sensitive Posts. In: 2022 IEEE 22nd International Conference on Software Quality, Reliability, and Security Companion (QRS-C), IEEE, pp 8–15. https://doi.org/10.1109/QRS-C57518.2022.00012, https://ieeexplore.ieee.org/document/10076960/

El Ouirdi M, Pais I, Segers J, El Ouirdi A (2016) The relationship between recruiter characteristics and applicant assessment on social media. Comput. Hum. Behav. 62:415–422. https://doi.org/10.1016/j.chb.2016.04.012, https://www.sciencedirect.com/science/article/pii/S0747563216302771

Fiesler C, Dye M, Feuston JL, Hiruncharoenvate C, Hutto CJ, Morrison S, Khanipour Roshan P, Pavalanathan U, Bruckman AS, De Choudhury M et al (2017) What (or who) is public? privacy settings and social media content sharing. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing, pp 567–580

Ford D, Behroozi M, Serebrenik A, Parnin C (2019) Beyond the Code Itself: How Programmers Really Look at Pull Requests. In: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS), IEEE, pp 51–60. https://doi.org/10.1109/ICSE-SEIS.2019.00014, https://ieeexplore.ieee.org/document/8797633/

Fukuyama F, Richman B, Goel A (2021) How to save democracy from technology: ending big tech's information monopoly. Foreign Aff 100:98

Garcia R, Treude C, La W (2023) Towards Understanding the Open Source Interest in Gender-Related GitHub Projects. http://arxiv.org/abs/2303.09727,

Gerber N, Gerber P, Volkamer M (2018) Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. Comput. Secur. 77:226–261. https://doi.org/10.1016/j.cose.2018.04.002, https://www.sciencedirect.com/science/article/pii/S0167404818303031

Gill AJ, Vasalou A, Papoutsi C, Joinson AN (2011) Privacy dictionary: a linguistic taxonomy of privacy for content analysis. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 3227–3236

Gousios G (2013) The GHTorent dataset and tool suite. In: 2013 10th Working Conference on Mining Software Repositories (MSR), pp. 233–236. https://doi.org/10.1109/MSR.2013.6624034

Guzman E, Azócar D, Li Y (2014) Sentiment analysis of commit comments in github: an empirical study. In: Proceedings of the 11th working conference on mining software repositories, pp 352–355

Hartigan JA, Wong MA (1979) Algorithm as 136: A k-means clustering algorithm. J Royal Stat Soc Ser C (Appl Stat) 28(1):100–108

Henderson KE (2019) They posted what? Recruiter use of social media for selection 48(4). https://doi.org/10.1016/j.orgdyn.2018.05.005

Henning A, Schulte L, Herbold S, Kulyk O, Mayer P (2023) Understanding issues related to personal data and data protection in open source projects on github. arXiv e-prints pp arXiv–2304

Hossin M, Sulaiman MN (2015) A review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manag Proc 5(2):1

Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146

Imtiaz N, Middleton J, Chakraborty J, Robson N, Bai G, Murphy-Hill E (2019) Investigating the Effects of Gender Bias on GitHub. In: 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE), pp 700–711. https://doi.org/10.1109/ICSE.2019.00079, https://ieeexplore.ieee.org/abstract/document/8812110?casa_token=LUfYyiYCTGMAAAAA:mDF9o-uDnunu01ee2s5rBcSUUdhApw4mNl6K92dbHz7CXvNuZolV0P4I-ZhqPrwa3nkQdsM

Inan H, Upasani K, Chi J, Rungta R, Iyer K, Mao Y, Tontchev M, Hu Q, Fuller B, Testuggine D, Khabsa M (2023) Llama guard: Llm-based input-output safeguard for human-ai conversations. 2312:06674

Inverardi P, Migliarini P, Palmiero M (2023) Systematic review on privacy categorisation. Comput Sci Rev 49

Iyer RN, Yun SA, Nagappan M, Hoey J (2019) Effects of Personality Traits on Pull Request Acceptance 47(11):2632–2643

Jay R (2000) Uk data protection act 1998 - the human rights context. Int Rev Law Comput Technol 14(3):385–395. https://doi.org/10.1080/713673366, https://doi.org/10.1080/713673366

Kanampiu M, Anwar M (2019) Privacy preferences vs. privacy settings: An exploratory facebook study. In: Advances in Human Factors in Cybersecurity: Proceedings of the AHFE 2018 International Conference on Human Factors in Cybersecurity, July 21-25, 2018, Loews Sapphire Falls Resort at Universal Studios, Orlando, Florida, USA 9, Springer, pp 116–126

Keküllüoglu D, Magdy W, Vaniea K (2020) Analysing privacy leakage of life events on twitter. In: Proceedings of the 12th ACM Conference on Web Science, pp 287–294

Khalajzadeh H, Shahin M, Obie HO, Grundy J (2022a) How are diverse end-user humancentric issues discussed on github? Association for Computing Machinery, New York, NY, USA, ICSE-SEIS '22, p 79–89. https://doi.org/10.1145/3510458.3513014

Khalajzadeh H, Shahin M, Obie HO, Grundy J (2022b) How are diverse end-user humancentric issues discussed on github? In: Proceedings of the 2022 ACM/IEEE 44th International Conference on Software Engineering: Software Engineering in Society, pp 79–89

King RS (2015) Cluster analysis and data mining: An introduction. Mercury Learn Inf

Kokolakis S (2017) Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. Comput Secur 64:122–134

Liu B, Andersen MS, Schaub F, Almuhimedi H, Zhang S, Sadeh NM, Agarwal Y, Acquisti A (2016) Follow my recommendations: A personalized privacy assistant for mobile app permissions. In: Twelfth Symposium on Usable Privacy and Security, SOUPS 2016, Denver, CO, USA, June 22-24, 2016, USENIX Association, pp 27–41. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/liu

Lustgarten SD, Garrison YL, Sinnard MT, Flynn AW (2020) Digit Priv Ment Healthc Curr Issues Recomm Technol Use 36:25–31. https://doi.org/10.1016/j.copsyc.2020.03.012, https://www.sciencedirect.com/science/article/pii/S2352250X20300415

Lu J, Yu L, Li X, Yang L, Zuo C (2023) Llama-reviewer: Advancing code review automation with large language models through parameter-efficient fine-tuning. In: 2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE), pp 647–658. https://doi.org/10.1109/ISSRE59848.2023.00026

Matz SC, Appel RE, Kosinski M (2020) Priv Age Psychol Target 31:116–121. https://doi.org/10.1016/j.copsyc.2019.08.010, https://www.sciencedirect.com/science/article/pii/S2352250X19301332

Meli M, McNiece MR, Reaves B (2019) How Bad Can It Git? Characterizing Secret Leakage in Public GitHub Repositories. In: Proceedings 2019 Network and Distributed System Security Symposium, Internet Society. https://doi.org/10.14722/ndss.2019.23418, https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_04B-3_Meli_paper.pdf

Migliarini P, Scoccia GL, Autili M, Inverardi P (2020) On the elicitation of privacy and ethics preferences of mobile users. In: Proceedings of the IEEE/ACM 7th International Conference on Mobile Software Engineering and Systems, pp 132–136

Miller T (2010) Surveillance: The “Digital Trail of Breadcrumbs" 2(1):9–14. https://doi.org/10.3384/cu.2000.1525.10219, https://cultureunbound.ep.liu.se/article/view/1913

Miller C, Cohen S, Klug D, Vasilescu B, Kästner C (2022b) “did you miss my comment or what?": Understanding toxicity in open source discussions. In: Proceedings of the 44th International Conference on Software Engineering, Association for Computing Machinery, New York, NY, USA, ICSE ’22, pp 710–722. https://doi.org/10.1145/3510003.3510111, https://doi.org/10.1145/3510003.3510111

Miller C, Cohen S, Klug D, Vasilescu B, KaUstner C (2022a) "Did you miss my comment or what?": Understanding toxicity in open source discussions. In: Proceedings of the 44th International Conference on Software Engineering, Association for Computing Machinery, ICSE ’22, pp 710–722. https://doi.org/10.1145/3510003.3510111, https://dl.acm.org/doi/10.1145/3510003.3510111

Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, Agirre E, Heintz I, Roth D (2023) Recent advances in natural language processing via large pre-trained language models: A survey. ACM Comput Surv 56(2). https://doi.org/10.1145/3605943, https://doi.org/10.1145/3605943

Nguyen TT, Wilson C, Dalins J (2023) Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts. http://arxiv.org/abs/2308.14683,

Niu L, Mirza S, Maradni Z, Pöpper C (2023) CodexLeaks: Privacy leaks from code generation language models in GitHub copilot. In: 32nd USENIX Security Symposium (USENIX Security 23), pp 2133–2150

Pardau SL (2018) The california consumer privacy act: Towards a european-style privacy regime in the united states. J Tech L & Pol’y 23:68

Peiretti F, Pensa RG (2023) Detection of Privacy-Harming Social Media Posts in Italian. In: Arief B, Monreale A, Sirivianos M, Li S (eds) Security and Privacy in Social Networks and Big Data, Springer Nature, Lecture Notes in Computer Science, pp 203–223. https://doi.org/10.1007/978-981-99-5177-2_12

Raman N, Cao M, Tsvetkov Y, Kästner C, Vasilescu B (2020) Stress and burnout in open source: Toward finding, understanding, and mitigating unhealthy interactions. In: Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results, Association for Computing Machinery, ICSE-NIER ’20, pp 57–60. https://doi.org/10.1145/3377816.3381732, https://dl.acm.org/doi/10.1145/3377816.3381732

Robillard MP, Walker RJ, Zimmermann T (2010) Recommendation systems for software engineering. IEEE Softw 27(4):80–86. https://doi.org/10.1109/MS.2009.161, https://doi.org/10.1109/MS.2009.161

Sajadi A, Damevski K, Chatterjee P (2023) Interpersonal Trust in OSS: Exploring Dimensions of Trust in GitHub Pull Requests. In: 2023 IEEE/ACM 45th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER), pp 19–24. https://doi.org/10.1109/ICSE-NIER58687.2023.00010, https://ieeexplore.ieee.org/abstract/document/10173872?casa_token=hBC8Zu0afC0AAAAA:-ozSXKFFvF78i5L9Srsw6nNuHzZYi5qzPKtg5MyrZsEPiKKnVEFtgnhUiP9Tulnkh-EapOo-OmE

Sanchez OR, Torre I, He Y, Knijnenburg BP (2020) A recommendation approach for user privacy preferences in the fitness domain. User Model User-Adap Inter 30:513–565

Solove DJ (2021) The myth of the privacy paradox. Geo Wash L Rev 89:1

Stretton T, Aaron L (2015) Dangers Our Trail Digit Breadcrumbs 2015(1):13–15. https://doi.org/10.1016/S1361-3723(15)70006-0, https://www.sciencedirect.com/science/article/pii/S1361372315700060

Syakur M, Khotimah B, Rochman E, Satoto BD (2018) Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In: IOP conference series: materials science and engineering, vol 336, p 012017. IOP Publishing

Tadesse MM, Lin H, Xu B, Yang L (2019) Detection of Depression-Related Posts in Reddit Social Media. Forum 7:44883–44893. https://doi.org/10.1109/ACCESS.2019.2909180, https://ieeexplore.ieee.org/abstract/document/8681445

Tahaei M, Vaniea K, Saphra N (2020) Understanding privacy-related questions on stack overflow. In: Proceedings of the 2020 CHI conference on human factors in computing systems, pp 1–14

Tang R, Han X, Jiang X, Hu X (2023) Does Synthetic Data Generation of LLMs Help Clinical Text Mining?

Terrell J, Kofink A, Middleton J, Rainear C, Murphy-Hill E, Parnin C, Stallings J (2017) Gender differences and bias in open source: Pull request acceptance of women versus men. PeerJ Comput Sci 3

Timoshenko A, Hauser JR (2019) Identifying Customer Needs from User-Generated Content 38(1):1–20. https://doi.org/10.1287/mksc.2018.1123, https://pubsonline.informs.org/doi/abs/10.1287/mksc.2018.1123

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Roziére B, Goyal N, Hambro E, Azhar F et al (2023) Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971

Umar P, Squicciarini A, Rajtmajer S (2019) Detection and Analysis of Self-Disclosure in Online News Commentaries. In: The World Wide Web Conference, ACM, pp 3272–3278. https://doi.org/10.1145/3308558.3313669, https://dl.acm.org/doi/10.1145/3308558.3313669

Vasalou A, Gill A, Mazanderani F, Papoutsi C, Joinson A (2011) Privacy dictionary: A new resource for the automated content analysis of privacy 62(11):2095–2105. https://doi.org/10.1002/asi.21610

Vasilescu B, Filkov V, Serebrenik A (2015a) Perceptions of Diversity on Git Hub: A User Survey. In: 2015 IEEE/ACM 8th International Workshop on Cooperative and Human Aspects of Software Engineering, pp 50–56. https://doi.org/10.1109/CHASE.2015.14, https://ieeexplore.ieee.org/abstract/document/7166088

Vasilescu B, Posnett D, Ray B, van den Brand MG, Serebrenik A, Devanbu P, Filkov V (2015b) Gender and tenure diversity in github teams. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp 3789–3798

Vasilescu B, Serebrenik A, Filkov V (2015c) A Data Set for Socia Diversity Studies of GitHub Teams. In: 2015 IEEE/ACM 12th Working Conference on Mining Software Repositories, pp 514–517. https://doi.org/10.1109/MSR.2015.77

Voigt P, Von dem Bussche A (2017) The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed, Cham: Springer International Publishing 10(3152676):10–5555

Wang J, Zhang X, Chen L, Xie X (2022) Personalizing label prediction for github issues. Inf Softw Technol 145:106845

Warrens MJ (2015) Five ways to look at cohen's kappa. J Psychol & Psychother 5

Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y (2024) A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. High-Confidence Comput. 4(2):100211. https://doi.org/10.1016/j.hcc.2024.100211, https://www.sciencedirect.com/science/article/pii/S266729522400014X

Zhang S, Feng Y, Bauer L, Cranor LF, Das A, Sadeh N (2021) "did you know this camera tracks your mood?": Understanding privacy expectations and preferences in the age of video analytics. Proc Priv Enhancing Technol 2021(2)

Zhang S, Feng Y, Das A, Bauer L, Cranor LF, Sadeh N (2020) Understanding people's privacy attitudes towards video analytics technologies. Proceedings of the FTC PrivacyCon, pp 1–18