Understanding Developers Privacy Concerns Through Reddit Thread Analysis

Jonathan Parsons[1,*,†], Michael Schrider[1,†], Oyebanjo Ogunlela[1] and Sepideh Ghanavati[1]

[1]The University of Maine, Orono, ME 04469

Abstract

With the growing global emphasis on regulating the protection of personal information and increasing user expectation of the same, developing with privacy in mind is becoming ever more important. In this paper, we study the concerns, questions, and solutions developers discuss on Reddit forums to enhance our understanding of their perceptions and challenges while developing applications in the current privacy-focused world. We perform various forms of Natural Language Processing (NLP) on 437,317 threads from subreddits such as r/webdev, r/androiddev, and r/iOSProgramming to identify both common points of discussion and how these points change over time as new regulations are passed around the globe. Our results show that there are common trends in privacy topics among the different subreddits while the frequency of those topics differs between web and mobile applications.

Keywords

Reddit, Developers, Privacy, Application development, Privacy policy, Natural language processing

1. Introduction

As the world is continuously advancing in ever-growing connected ways, software developers and requirements analysts are required to implement privacy-preserving solutions to protect users' privacy in their applications. On the other hand, users are becoming more conscious about how their information is collected and used by various organizations. These parallel growths resulted in the creation of several privacy-focused regulations, such as the General Data Protection Regulation (GDPR) [1] and California Consumer Protection Act (CCPA) [2].

To ensure compliance with these regulations, systematic privacy by design approaches and tools need to become the norm. Without proper privacy education, tools, and guidelines, developers look into forums such as Stack Overflow [3], or Reddit [4] to find solutions to their

In: A. Ferrari, B. Penzenstadler, I. Hadar, S. Oyedeji, S. Abualhaija, A. Vogelsang, G. Deshpande, A. Rachmann, J. Gulden, A. Wohlgemuth, A. Hess, S. Fricker, R. Guizzardi, J. Horkoff, A. Perini, A. Susi, O. Karras, A. Moreira, F. Dalpiaz, P. Spoletini, D. Amyot. Joint Proceedings of REFSQ-2023 Workshops, Doctoral Symposium, Posters & Tools Track, and Journal Early Feedback Track. Co-located with REFSQ 2023. Barcelona, Catalunya, Spain, April 17, 2023.

[*]Corresponding author.

[†]These authors contributed equally.

jonathan.parsons1@maine.edu (J. Parsons); michael.schrider@maine.edu (M. Schrider); oyebanjo.ogunlela@maine.edu (O. Ogunlela); sepideh.ghanavati@maine.edu (S. Ghanavati)

https://www.sepidehghanavati.com/ (S. Ghanavati)

0000-0003-2778-8136 (J. Parsons); 0000-0002-3725-7438 (M. Schrider); 0000-0000-0000-0000 (O. Ogunlela); 0000-0001-7972-667X (S. Ghanavati)

privacy-related questions. Understanding the types of privacy-focused questions asked on these forums and developers' challenges helps better tailor the tools and approaches to their needs.

In addition to surveys [5], in recent years, some work focused on evaluating privacy-related questions on Stack Overflow (SO) and Reddit [6, 7, 8]. Li et al. [8] conducted an analysis of Android developers' privacy concerns on r/anddroidev subreddit and discovered that privacy appears to be underrepresented. Tahaei et al. [6, 7] evaluated SO and show that conversations are mostly focused on compliance with regulations, often citing official documents. Li et al. [9] used Reddit [10] to identify the narratives driving users' privacy concerns rather than developers and showed that understanding the users' concerns could drive developers' concerns, as well.

In this paper, we will extend Li et al. [8] approach by evaluating other similar subreddits, /r/iOSprogramming and /r/webdev, and conduct a comparative analysis on the types of privacy questions asked based on privacy requirements imposed by these three frameworks (i.e., Android, iOS, and Web) and the role of privacy regulations on the questions. We also examine the sentiment regarding privacy, comparing that amongst various subreddit communities to get a better idea about how developers actually view privacy, not just how they comply with it. Our analysis of 437,317 threads shows that there are differences in the most frequent topics based on mobile and web app development. We also observe that the regulations such as GDPR and CCPA impacted the topics' trend, however, we could not conclude a change in the overall sentiment. Our research questions which we will answer in this paper are as follows:

- **RQ1:** What are the major privacy concerns on developer forums?
- **RQ2:** What is the overall sentiment of developing privacy requirements in the studied communities? How does it differ by subreddits?
- **RQ3:** How do regulations such as GDPR and CCPA influence RQ1 and RQ2?

2. Related Works

In recent years, several works focused on evaluating and understanding the privacy challenges of developers and their approach to implementing privacy requirements. Some studies evaluate developers' privacy behaviors through questionnaires or surveys [11, 12, 13] and show that developers generally see privacy as a burden and afterthought and are not familiar with basic privacy concepts [14, 5]. Other research [6, 7, 8, 15] evaluate developers' popular forums such as Stack Overflow (SO) [3] and Reddit [4] to identify the topics of developers' questions.

Proferes et al. [10] analyzed 727 publications between 2010 and 2020 and identified a substantial increase in the usage of Reddit as a data collection medium by disciplines ranging from computer science to social sciences. Iqbal et al. [16] evaluated the data from the top 10 mobile and desktop apps' subreddits and identified that 54% of the post included useful information such as bug reports or feature requests and could be used for requirements elicitation.

Analysis of 4,957 Reddit comments in 180 security- and privacy-related discussions from /r/homeautomation show that users' concerns are context-dependent and their attitude towards privacy and security can change during the different phases of adoption of smart home devices [17]. Li et al. [8] analyzed 207 discussions on r/androiddev subreddit to identify how developers discuss personal data protection. Their findings indicate that privacy concerns are not discussed often on developer's forums and developers shy away from discussions relating to privacy concerning plan and execution trials. However, they posit that they seem externally motivated by new demands for privacy emerging from privacy-focused regulations. Tahaei et al. [6, 7, 18] studied 315 privacy questions on SO and identified that the introduction of Google and Apple privacy labels resulted in an increase in the number of privacy-related questions on SO [6]. In another study [7], they explored the types of advice given by developers on privacy issues and compared them with Hoepman's approach [19]. They identified 148 pieces of advice focused on regulations and confidentiality, including 'inform,' 'hide,' 'control,' and 'minimize'.

In this paper, we extend the current approaches [7, 8, 11, 18] in two ways: (a) the number of subreddits to review and (b) the scope at which they are analyzed. We extend Li et al. [8] by adding r/iOSprogramming and r/webdev to r/androiddev to get a more general idea of developers' privacy concerns. These subreddits constitute the majority of the current software development. We also conduct *sentiment* analysis on the discussions similar to [20, 21, 22], and evaluate how those may change between the different communities to gain insight into the emotions behind the developers when it comes to privacy and the surrounding policy.

3. Methodologies

In this section, we describe our methodology for collecting and creating the Reddit dataset (i.e., Section 3.1) and then explain the process of analyzing the data (i.e., Section 3.2).

3.1. Gathering Reddit Data

To gather data, we utilize Pushshift Multithread API Wrapper [23, 24] which allows for granular information and provides the ability to multithread, in compare to other work [9, 20, 25].

The current related work provides insights regarding *what* and *how* of extracting the information. Li et al. [8] focus on /r/androiddev, which includes 203k members (top 1% of subreddits by size) [26] and is an active community with the discussions limited to high-level android app design. We extend their effort to two other similar subreddits, /r/iOSprogramming (120k members) [27] and /r/webdev (1.4mil members) [28] to broaden the scope of the analysis of developers' privacy concerns across various development platforms. In total, we pulled 100,040 submissions from r/androiddev, 55,553 submissions from r/iOSprogramming, and 281,724 submissions from r/webdev along with all associated comments, from January 2014 through November 2022. We chose these subreddits specifically due to their relative similarity in being platform-specific, developer-centric forums.

3.2. Processing and Analyzing Reddit Data

After collecting the data, we need to identify privacy-related submissions. A common way to do this is by filtering posts and discussions that contain the term "privacy" [6, 7], or other similar terms that are often used in discussions around privacy such as "GDPR", "CCPA", "mac address", "location", etc [8, 9, 21]. In Table 1, we augmented the terms identified by Li et al. [8] (i.e., Unique Identifier, Photo, and Video, Audio, and Location) with general privacy and privacy regulation terms (i.e., Privacy category in Table 1). We then used these keywords to create

Table 1
A List of the Privacy Terms

Data Category	Privacy Keywords
Privacy	private, privacy, gdpr, general data protection regulation, ccpa, ccpr, califronia consumer privacy act, California consumer privacy regulation
1 Unique Identifier	first name, last name, real name, identification number, id number, social security number, ssn, license number, passport number, screen name, user name, account name, user id, username, userid, online identifier, imei, device serial number, advertising id, android id, ssaid, mac address, imsi, instance id, guid, internet protocol address, ip address, email address, telephone number, phone number, line1
Photo and Video	video recording, camera, gallery
Audio	audio recording, microphone, voice
Location	voice, location, physical street address, home address, physical address, street name, city name, postal address

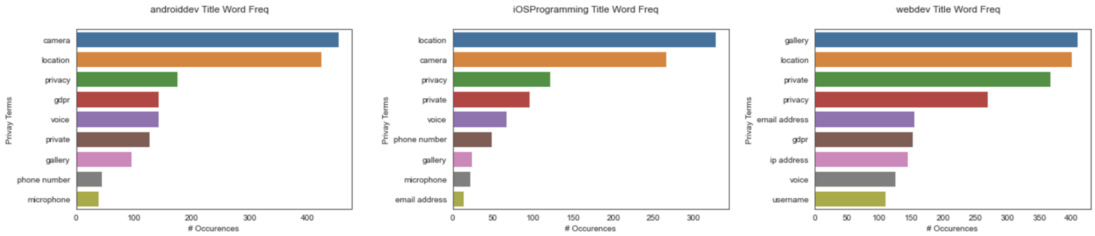
a privacy-related dataset before performing our analysis. We also preprocessed the data by removing stopwords, stemming, and lemmatizing text to prepare for our analysis [29].

To answer RQ1, we conducted simple phrase frequency analysis and then identified the topics of each submission by leveraging Latent Dirichlet Allocation (LDA) [30] similar to other work [29]. We first narrowed down our dataset to privacy questions via an Adaptive Boosting (AdaBoost) model to classify posts in our dataset as questions. The AdaBoost classifier contained 25 decision stumps and was trained on a combination of subsets of SQuAD and SPAADIA datasets which include phrases and sentences labeled as either statements or questions [31, 32, 33]. We transformed the training data into matrices of token counts and validated the classifier against 200 randomly sampled posts. We manually classified posts from AndroidDev containing 93 (46.5%) questions and 107 (53.5%) statements. The classification was performed by a single team member and verified by another; the agreed-upon method was to read both the title and body, and if either had a question from the redditer to the community, then it was classified as a question; framing and rhetorical questions were not considered questions. Against the validation set, our classification achieved 71% accuracy. Through LDA analysis, we generated ten topics of four words each for posts Pre/Post GDPR and Pre/Post CCPA.

To address RQ2, we use qualitative metrics to evaluate both the discussions and posts [10]. Qualitatively, we perform sentiment analysis leveraging Natural Language Toolkit (NLTK) approaches [20, 21, 34].

We answer RQ3 by applying the same term frequency analysis and LDA topic analysis used for RQ1 against datasets filtered to pre- and post-regulations and trending the RQ2 sentiment analysis to examine if there is any change due to the introduction of GDPR (April 2016) and CCPA (June 2018).[2]

[2] The data, models, and analysis are: https://github.com/mschrider/PEP_Privacy_Dev_Forum_Analysis/.

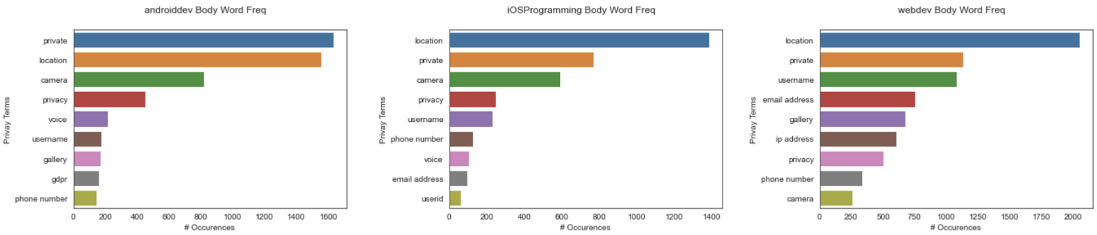


(a) Android Title Word Frequency

(b) iOS Title Word Frequency

(c) Web Title Word Frequency

Figure 1: Comparison of Title Text Frequency



(a) Android Body Word Frequency

(b) iOS Body Word Frequency

(c) Web Body Word Frequency

Figure 2: Comparison of Body Text Frequency

4. Results

In the first step, we conducted a simple word counting analysis to identify how often privacy-related words appear in the initial posts of r/androiddev, r/iOSProgramming, and r/webdev forums. We noticed that typical permission requests, such as “location”, “camera”, “gallery”, “microphone” and “voice” are the more prevalent words, though they still make up a small portion of the overall posts. Additionally, we observed that the most common word, *location*, appears nearly twice as often as the second common one, *camera*. With more detailed analysis, we noticed that while *location* and *camera* are privacy-related words, they are more often referred to for non-privacy reasons; such as the best methods for having an app interact with a camera. After this high-level evaluation, we delve into our research questions.

RQ1 - What are the major privacy concerns on developer forums?

The initial phrase frequency analysis indicates that “location” and “camera” are the main privacy-related terms referenced in the titles and bodies of posts. Meanwhile /r/webdev prioritizes “location” but replaces “camera” with “gallery” while “email address” and “username” are much higher up in the list (See Fig. 1 and Fig. 2). This data shows the inherent difference between the primary functions of web-based and mobile-based applications.

Based on the trends above, we narrowed down the questions to those containing “location” or “camera” for further analysis. We observe that most questions are not in fact privacy-related;

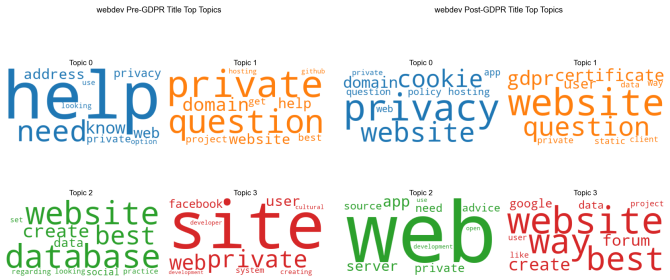


Figure 3: webdev Top Topics Pre (left) and Post (right) GDPR.

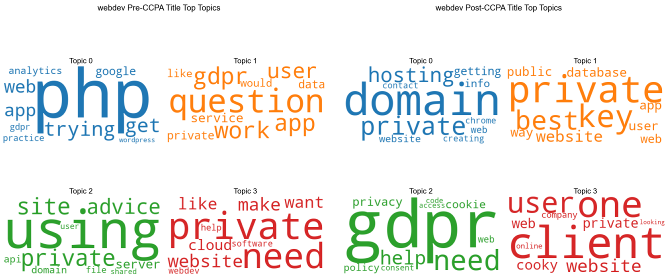


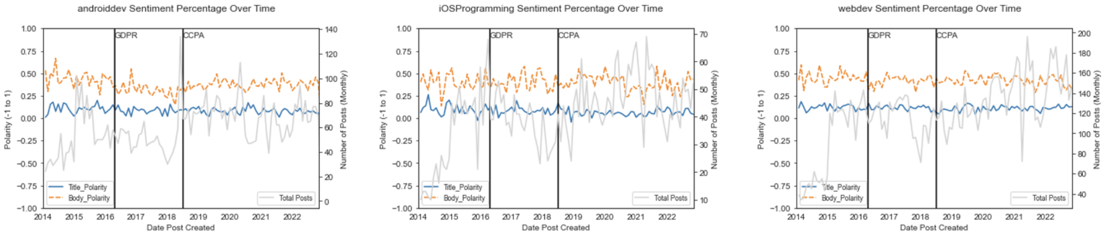
Figure 4: webdev Top Topics Pre (left) and Post (right) CCPA.

though for *location*, one of the most upvoted questions is “GDPR - What all do I need to do?”. In general, we identified that the questions that were actually privacy-related were centered around asking for and granting permission to apps for capabilities that have impacts on privacy.

Our LDA analysis shows common topics among privacy questions. Fig. 3 and Fig. 4 include the top four webdev topics for pre- and post-GDPR/CCPA. We observed that ‘gdpr’ is one of the top trending topics from its expected non-existence pre-GDPR. Topics with variations of the word policy, consent, or cookie increased in frequency post-GDPR/CCPA but privacy-related topics appear more frequently post-CCPA. There is an overlap in post-CCPA with post-GDPR data; thus, it may be hard to distinguish the effects of each regulation in isolation.

RQ2 - What is the overall sentiment of developing privacy in the studied communities? Does it differ by subreddit?

Overall, all subreddits show similar sentiment profiles, with titles generally tending toward neutral and the main bodies of text tending toward positive. The results in Fig. 5 show the trends over time, with markers for where GDPR and CCPA were introduced. While all subreddits show undulations, there is no concrete evidence of trends in sentiment shifting one way or another.



(a) Android Sentiment Over Time

(b) iOS Sentiment Over Time

(c) Web Sentiment Over Time

Figure 5: Sentiment over time for androiddev, iOSProgramming, and webdev subreddits.

RQ3 - How do regulations such as GDPR and CCPA influence RQ1 and RQ2?

Topics and terms show a significant change due to GDPR but to a lesser extent CCPA. As shown in Fig. 3 and Fig. 4, the overall distribution of topics/terms changed noticeably after both regulations were enacted. However, despite the changing topics, we did not observe conclusive evidence of a change in sentiment post-GDPR/CCPA. All three subreddits showed very steady sentiment, with roughly similar noise values, across the entire timeline represented in the data.

5. Discussion

Similarities between /r/iOSProgramming and /r/androiddev are expected since both deal with mobile development. Mobile privacy concerns such as camera, microphone, and location permissions dominate the discussions while on /r/webdev, main focus is on cookies and websites.

Understanding the trends and commonalities regarding developers' privacy concerns is important to platform owners (e.g., Apple, Google, etc.), regulators, and requirements analysts. Platform owners should provide support to mitigate and resolve developers' privacy questions. Regulations drive developers' concerns; therefore, regulators should study the impact of their regulations on developer behaviors. With this information, requirements analysts could focus on developing approaches to elicit and model privacy requirements from regulations and automate the compliance process.

This study was done solely on Reddit, which may introduce an inherent selection bias; for example, due to focusing on issues brought up by developers active in their online communities. While we extended previous research [8], it still leaves out significant areas of research such as conducting interviews, surveys, or analysis of other forums to gain insights.

There are limitations to our NLP analysis. The AdaBoost question classifier does not consider relations between tokens, parts of speech utilized in Reddit posts, or other potential features which leads to improper classification of rhetorical questions. For example, a post with: "Heard about a cool job posting? Let people know!" was misclassified. The sentiment analysis is also generic since the corpus used to generate the polarity scores, nltk.sentiment.vader, is specifically designed to be used with social media posts. While Reddit is a social media platform, a more directed corpus around software development most likely provides more accurate results. In the future, we will explore approaches proposed in [22, 35] to conduct a more detailed sentiment analysis. Lastly, our data collection for privacy-related questions is based on keywords search which may result in missing a large number of privacy content. We propose to use Natural Language Inference (NLI) approaches to extract a larger pool of data, similar to [36].

6. Conclusion and Future Work

In this paper, we looked into the developers' concerns and questions on Reddit forums to gain a better understanding of their attitudes toward privacy and the challenges they face when developing applications. We examined 437,317 threads from the subreddits r/webdev, r/androiddev, and r/iOSProgramming to determine the most frequently discussed topics as well as how the sentiments around these topics have evolved in response to GDPR and CCPA. Through a combination of word frequency, topic clustering, and question classifying, we observed that a large number of questions are related to requesting permissions from users, such as camera and location, or complying with the various regulations, particularly the GDPR. Additionally, we explored the emotions around dealing with privacy and found that there is a general neutral-to-positive sentiment around it across all types of development reviewed.

In the future, we plan to extend our effort to other developers' forums and improve our classification task. We will also leverage NLI to extract more privacy-related concepts.

References

[1] The EU, The General Data Protection Regulation, http://www.eugdpr.org/, accessed, 2023.

[2] Government of California, California Consumer Privacy Act (CCPA), https://oag.ca.gov/privacy/ccpa, accessed 2023).

[3] Stack Overflow, https://stackoverflow.com/, (accessed 2023).

[4] Reddit, https://www.reddit.com/r/programming/, (accessed 2023).

[5] I. Hadar, T. Hasson, O. Ayalon, E. Toch, M. Birnhack, S. Sherman, A. Balissa, Privacy by designers: Software developers' privacy mindset, Journal of Empirical Software Engineering 23 (2018) 259–289. doi:10.1007/s10664-017-9517-1.

[6] M. Tahaei, K. Vaniea, N. Saphra, Understanding privacy-related questions on stack overflow, in: Proc. of the 2020 CHI Conf. on Human Factors in Computing Systems, 2020, p. 1–14.

[7] M. Tahaei, T. Li, K. Vaniea, Understanding privacy-related advice on stack overflow, Proceedings on Privacy Enhancing Technologies 2022 (2022) 114–131.

[8] T. Li, E. Louie, L. Dabbish, J. I. Hong, How developers talk about personal data and what it means for user privacy: A case study of a developer forum on reddit, Proc. ACM Hum.-Comput. Interact. 4 (2021).

[9] Z. S. Li, M. Sihag, N. N. Arony, J. B. Junior, T. Phan, N. Ernst, D. Damian, Narratives: the unforeseen influencer of privacy concerns (2022). URL: https://arxiv.org/abs/2206.05600.

[10] N. Proferes, N. Jones, S. Gilbert, C. Fiesler, M. Zimmer, Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics, Social Media+ Society 7 (2021).

[11] M. Tahaei, K. Vaniea, "developers are responsible": What ad networks tell developers about privacy, in: Extended Abstracts CHI'21, 2021, pp. 1–11.

[12] A. Ekambaranathan, J. Zhao, M. Van Kleek, “money makes the world go around”: Identifying barriers to better privacy in children’s apps from developers’ perspectives, in: Proc. of the 2021 CHI Conf. on Human Factors in Computing Systems, 2021, pp. 1–15.

[13] S. Spiekermann, J. Korunovska, M. Langheinrich, Inside the organization: Why privacy and security engineering is a challenge for engineers, Proc. of the IEEE 107 (2018) 600–615.

[14] N. Alomar, S. Egelman, J. L. Fischer, Developers say the darnedest things: Privacy compliance processes followed by developers of child-directed apps, Proc. on PETS (2022).

[15] D. Greene, K. Shilton, Platform privacies: Governance, collaboration, and the different meanings of “privacy” in ios and android development, New Media & Society (2018).

[16] T. Iqbal, M. Khan, K. Taveter, N. Seyff, Mining reddit as a new source for software requirements, in: IEEE 29th Int. Requirements Engineering Conf., 2021, pp. 128–138.

[17] J. Li, K. Sun, B. Huff, A. Bierley, Y. Kim, F. Schaub, K. Fawaz, “it’s up to the consumer to be smart”: Understanding the security and privacy attitudes of smart home users on reddit, in: IEEE Symp. on Security and Privacy (SP) (SP), 2023, pp. 380–396.

[18] M. Tahaei, J. Bernd, A. Rashid, Privacy, permissions, and the health app ecosystem: A stack overflow exploration, in: The European Symp. on Usable Security, 2022, p. 117–130.

[19] J.-H. Hoepman, Privacy design strategies (the little blue book) (2018).

[20] B. Neo, Sentiment analysis on reddit tech news with python (2021). URL: https://tinyurl.com/yck5u7ke.

[21] D. Pletea, B. Vasilescu, A. Serebrenik, Security and emotion: Sentiment analysis of security discussions on github, in: The 11th Working Conf. on MSR, ACM, 2014, p. 348–351.

[22] M. Herrmann, M. Obaidi, L. Chazette, J. Klünder, On the subjectivity of emotions in software projects: How reliable are pre-labeled data sets for sentiment analysis?, Journal of Systems and Software 193 (2022).

[23] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, J. Blackburn, The pushshift reddit dataset (2020). URL: https://arxiv.org/abs/2001.08435. doi:10.48550/ARXIV.2001.08435.

[24] M. Podolak, Pmaw: Pushshift multithread api wrapper, 2022. URL: https://pypi.org/project/pmaw/.

[25] A. Amaya, R. Bach, F. Keusch, F. Kreuter, New data sources in social science research: Things to know before working with reddit data, Social science computer review (2021).

[26] Reddit, Developing android apps, 2022. URL: https://www.reddit.com/r/androiddev/.

[27] Reddit, ios programming, 2022. URL: https://www.reddit.com/r/iOSProgramming/.

[28] Reddit, webdev: reddit for web developers, 2022. URL: https://www.reddit.com/r/webdev/.

[29] G. L. Scoccia, P. Migliarini, M. Autili, Challenges in developing desktop web apps: a study of stack overflow and github, in: IEEE/ACM 18th Int. Conf. on MSR, 2021, pp. 271–282.

[30] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, JMLR 3 (2003) 993–1022.

[31] M. Weisser, Corpus data, 2020. URL: https://martinweisser.org/.

[32] P. Rajpurkar, R. Jia, P. Liang, Know what you don’t know: Unanswerable questions for squad, CoRR (2018). URL: http://arxiv.org/abs/1806.03822. arXiv:1806.03822.

[33] S. Khan, Questions vs statements classification, 2021. URL: https://www.kaggle.com/datasets/shahrukhkhan/questions-vs-statementsclassificationdataset.

[34] N. Xue, Steven bird, evan klein and edward loper. natural language processing with python. o’reilly media, inc.2009., Natural Language Engineering 17 (2011) 419–424.

[35] N. Novielli, F. Calefato, D. Dongiovanni, D. Girardi, F. Lanubile, Can we use se-specific

sentiment analysis tools in a cross-platform setting?, in: Proceedings of the 17th Int. Conf. on MSR, MSR ’20, 2020, p. 158–168.

[36] H. Harkous, S. Peddinti, R. Khandelwal, A. Srivastava, N. Taft, Hark: A deep learning system for navigating privacy feedback at scale, in: IEEE Symp. on Sec. & Privacy, 2022.