

Hurdles for Developers in Cryptography

Mohammadreza Hazhirpasand
Oscar Nierstrasz
University of Bern
Bern, Switzerland

Mohammadhossein Shabani
Azad University
Rasht, Iran

Mohammad Ghafari
School of Computer Science
University of Auckland
Auckland, New Zealand

Abstract—Prior research has shown that cryptography is hard to use for developers. We aim to understand what cryptography issues developers face in practice. We clustered 91 954 cryptography-related questions on the Stack Overflow website, and manually analyzed a significant sample (*i.e.*, 383) of the questions to comprehend the crypto challenges developers commonly face in this domain. We found that either developers have a distinct lack of knowledge in understanding the fundamental concepts, *e.g.*, OpenSSL, public-key cryptography or password hashing, or the usability of crypto libraries undermined developer performance to correctly realize a crypto scenario. This is alarming and indicates the need for dedicated research to improve the design of crypto APIs.

Index Terms—Security, cryptography, developer issues

I. INTRODUCTION

Studies have shown that cryptography concepts are hard to understand for developers, and the complexity of crypto APIs has rendered their secure usage very difficult [1] [2]. There exist static analysis tools, but developers are reluctant to employ them due to a lack of familiarity, restrictions in organizational policies, or high rates of false positives [3], [4]. Researchers have recently developed new APIs to ease the adoption of cryptography [5], yet online Q&A forums are among the main information sources used to resolve developer issues.

Closer inspection of online forums such as Stack Overflow provides a shortcut to identifying the frequent challenges that developers face in this domain. Therefore, in this study, we address the following research question: *What types of crypto challenges do developers face in cryptography?* We extract the common problems that developers recently encounter when dealing with various areas of cryptography. The findings provide significant help for developers in general, and software team leaders, tutors and crypto library designers in particular, to raise their awareness of common misunderstandings, or to highlight areas with a steep learning curve.

Unlike other studies, we only focus on crypto-related challenges of developers. To cover various types of crypto-challenges, we need to identify different groups of questions that are similar in terms of context. Particularly, manual grouping of such a large number of questions (*i.e.*, 91 954) is a demanding task. We therefore used the Latent Dirichlet Allocation (LDA) generative statistical model, and found three main topics in 91 954 crypto-related posts on Stack Overflow. We then used stratified sampling to study 383 posts randomly from the three topics to identify the most common problematic

issues for developers. The results showed that developers commonly failed to implement a cryptographic scenario due to two reasons, namely the complexity of crypto APIs, and their lack of familiarity with fundamental concepts such as digital certificates, public-key cryptography, and hashing algorithms.

Our findings show that hurdles for developers in cryptography are not yet resolved, and due to its impact on security, this domain urgently needs dedicated research effort. We are conducting a survey with developers who actively helped the Stack Overflow community in this domain to understand potential remedies to this problem.

II. RELATED WORK

Sifat *et al.* investigated three online sources, *i.e.*, Crypto Stack Exchange, Security Stack Exchange, and Quora, to identify complications with respect to implementing security in data transmission [6]. Their findings suggest that the most discussed technique is transport layer security (TLS), and the Cross-Site Scripting (XSS) attack is the main concern of developers. In another study, Yang *et al.* conducted a large-scale analysis of security-related questions on Stack Overflow [7]. They identified five main categories, *i.e.*, web security, mobile security, cryptography, software security, and system security but they did not look into the challenges of each topic. A recent study conducted by Meng *et al.* has recognized the challenges of writing secure Java code on Stack Overflow [8]. Their examinations provide compelling evidence that security implications of coding options in Java, *e.g.*, CSRF tokens, are not well-perceived by a large number of developers. Nandi *et al.* conducted an empirical study on the frequent crypto obstacles with which Java developers commonly face [1]. They triangulated data from a survey, 100 randomly selected Java GitHub repositories, and the top 100 Java cryptography questions asked on Stack Overflow. Their analyses depicted nine main crypto topics, suggesting that developers face difficulties using cryptography. This issue has adversely affected developer performance and software security [9]. A recent study showed that developers blindly use the provided vulnerable code snippets found on Stack Overflow [10]. They mentioned that 15.4% of the 1.3 million Android applications contained security-related code snippets from Stack Overflow. The previous studies solely focused on security or crypto implications of a particular language or in general security-related concerns. In contrast, we specifically

analyzed crypto-related questions of any kind irrespective of any programming languages or particular part of cryptography.

III. METHODOLOGY

We first explain the data gathering procedure and then describe how we clean the data, and briefly introduce the LDA topic modeling.

A. Data Extraction

To collect crypto-related posts on Stack Overflow, we assumed that the attached tags to a question mainly reflect the question's topic. We first used the "cryptography" tag, *i.e.*, *base tag*, to fetch crypto-related posts, *i.e.*, 11 130 posts, with the help of the Data Explorer platform (Stack Exchange). We found 2 184 tags (*candidate tags*) that occurred in posts together with the "cryptography" tag. However, not all candidate tags were crypto-related *e.g.*, C#.

To find relevant posts with the base tag, we used two metrics to determine which of the candidate tags are exclusively related to the base tag. We introduced the first metric as *affinity* to determine the degree to which a candidate tag (T) is exclusively associated with the base tag (BT). For each T , we used the *posts with tags* function, for brevity `pwt()`, to calculate the number of posts whose tags contain both T and BT . We used `pwt()` to obtain the number of posts whose tags contain T . Given these two values, we compute $affinity(T, BT) = |pwt(T, BT)| / |pwt(T)|$, whose result ranges from zero to one.

The smaller the value of the first metric, the weaker the association between T and BT . For example, the "C++" and "encryption" tags each appeared 639 897 and 29 737 times respectively in the entire Stack Overflow. The "C++" tag appeared together with BT 540 times and "encryption" was used 3535 times with BT . The value of affinity for the "C++" tag is 0.0008 and 0.1188 for the "encryption" tag, values which demonstrate a strong affinity for "encryption" and BT .

However, higher values of affinity for some candidate tags do not necessarily indicate tags that are closely related to cryptography. For example, the "s60-3rd-edition" tag appeared once with the base tag and in total 11 times in Stack Overflow. The value of affinity for this candidate tag is 0.09, which is close to the value of the "encryption" tag, even though it appeared only once with the base tag. To resolve this issue, we introduced a second metric, *coverage*(T, BT) = $|pwt(T, BT)| / |pwt(BT)|$. The second metric indicates the coverage of the BT posts by T . As an example, the value (*i.e.*, 0.00008) of coverage for the "s60-3rd-edition" tag proves that the candidate tag does not exclusively cover the base tag while the "C++" tag covers 0.04 of the cryptography-related questions.

Two authors of this paper examined various combinations of thresholds for the two metrics, and manually reviewed the resulting tags. We noticed that the thresholds to collect *only crypto-related tags* from the candidate tags (*i.e.*, 2 184) are the ones above the affinity: 0.025 and coverage: 0.005. There are 40 crypto-related tags that fall within the selected threshold domain. The list of crypto-related tags as well as

their frequencies are available online.¹ Next, we again used Stack Exchange Data Explorer to extract posts containing each of the selected tags (*i.e.*, 40 tags) but not the base tag, and recorded them in CSV files, which are available online.²

B. Data clustering via Topic Modeling

We combined the title and body of a post in order to create a document. We removed duplicate post IDs in multiple CSV files, and finally obtained 91 954 unique documents, without considering when the posts were created. Evidently, each of the documents contained a large number of unnecessary text elements that could produce noise in the output of a topic modeling algorithm. We preprocessed the documents in the following steps: (1) we removed all the code blocks enclosed by the "<code>" tag, (2) we removed all the HTML elements with the help of the Beautiful Soup library,³ (3) we removed newlines and non-alphanumeric characters, (4) we used the NLTK package to eliminate English stop words from the documents, and finally (5) we used the Snowball stemmer to normalize the text by transforming words into their root forms, *e.g.*, playing converts to play. We found 269 795 stemmed words in total. Finally, we used the CountVectorizer class in Scikit-learn to transform the words into a vector of term/token counts to feed into a machine learning algorithm.

We used Scikit-learn,⁴ a popular machine learning library in Python that provides a range of supervised and unsupervised learning algorithms. Latent Dirichlet Allocation (LDA) is an unsupervised learning algorithm based on a generative probabilistic model that considers each topic as a set of words and each document as a set of topic probabilities [11]. LDA has been used to discover latent topics in documents in a large number of prior studies [12] [7] [13].

Before training a model, LDA requires a number of important parameters to be specified. LDA asks for a fixed *number of topics* and then maps all the documents to the topics. The *Alpha* parameter describes document-topic density, *i.e.*, higher alpha means documents consist of more topics, and generates a more precise topic distribution per document. The *Beta* parameter describes topic-word density, *i.e.*, higher beta means topics entail most of the words, and generates a more specific word distribution per topic.

The optimal values of hyperparameters cannot be directly estimated from the data, and, more importantly, the right choice of parameters considerably improves the performance of a machine learning model [14]. We therefore used the Grid-SearchCV function in Scikit-learn to perform hyperparameter tuning to generate candidates from an array of values for the three aforementioned parameters, *i.e.*, *Alpha*, *Beta*, and the *number of topics*. As research has shown that choosing the proper number of topics is not simple in a model, an iterative approach can be employed [15] to render various models with different numbers of topics, and choose the number of topics

¹http://185.94.98.132/~crypto/paper_data/tags.csv

²http://185.94.98.132/~crypto/paper_data/

³<https://www.crummy.com/software/BeautifulSoup/>

⁴<https://scikit-learn.org/>

for which the model has the least perplexity. Perplexity is a measure used to specify the statistical goodness of fit of a topic model [11]. We therefore specified the number of topics from 1 to 25. We also used the conditional hyperparameter tuning for Alpha, which means a hyperparameter may need to be tuned depending on the value of another hyperparameter [16]. We set $\alpha = 50 / \text{number of topics}$ and $\beta = 0.01$, following guidelines of previous research [17].

Optimizing for perplexity, however, may not always result in humanly interpretable topics [18]. To facilitate the manual interpretation of the topics, we used a popular visualization package, named pyLDAvis⁵, in Python. The two authors of this paper separately checked the resulting top keywords of the topics, *i.e.*, from 1 to 25, and the associated pyLDAvis visualizations to ensure that the given number of topics is semantically aligned with human judgment.

C. Data analysis

We computed the required sample size for 91 954 documents with a confidence level of 95% and a margin of error of 5%, which is 383 documents. We then used stratified sampling to divide the whole population into smaller groups, called strata. In this step, we considered each topic as one stratum, and randomly selected the documents proportionally from the different strata. We then used thematic analysis, a qualitative research method for finding topics in text [19], to extract the frequent topics from the documents. Two authors of the paper carefully reviewed the title, question body, and answer body of each document. Each author then improved the extracted topics by labeling the posts iteratively. We then calculated Cohen’s kappa, a commonly used measure of inter-rater agreement [20], between the two reviewers. Finally, the two reviewers compared their final labelling results, and re-analyzed the particular posts in a session where they disagreed in order to discuss and arrive at a consensus.

IV. RESULTS AND DISCUSSION

Our hyperparameter tuning demonstrated that the best number of topics is three. Similarly, after analyzing pyLDAvis’s visualizations and top keywords for 1 to 25 topics, the two reviewers also achieved a consensus on three as the number of topics. The pyLDAvis interactive visualization for the three topics is available online.⁶ The reviewers named the topics by considering the general themes of top keywords returned by LDA (See Table I). We determined that the first topic is about digital certificates and configuration issues, the second one is about programming issues concerning encryption and decryption, and the third concerns passwords/hashes and basic crypto-related algorithms. As an influential indicator of topic relevancy, we realized that the frequencies of the candidate tags used in the three topics are aligned with the general themes of the topics.⁷ For instance, we observed that the AES, DES, Encryption, and RSA tags are mostly used in

TABLE I
THE THREE TOPICS AND THEIR TOP KEYWORDS

Topic	Top keywords
Digital certificate and configuration problems	use, certif, file, server, key, openssl, client, work, tri, need, sign, user, error,applic, creat, code, secur, app, encrypt, ssl, store, instal, like, connect, problem, want, way, run, request
Programming issues	key, encrypt, use, decrypt, code, data, file, string, tri, public, work, ae, im, byte, need, java, generat, messag, encod, privat, rsa, cipher, algorithm, block, like, implement, error, problem, function, text
Password/hashes and basic crypto algorithms	hash, use, valu, password, function,like, array, string, need, code, number, key, want, way, store,data, tabl, im, salt, tri, differ, time, algorithm, work, md5, user, make, generat, object, implement

programming issues, the Hash, SHA, SHA256, MD5, XOR, and Salt tags are more frequent in the password/hash topic, and finally, the Digital-signature, Keystore, OpenSSL, Private-key, Public-key, Smartcard, and X509certificate tags are more common in the digital certificate topic.

With respect to stratified sampling, we considered the number of documents in each stratum (*i.e.*, each topic) as 139, 124, and 119 documents from the first topic to the third one respectively. The selected documents were created in the last 5 years on Stack Overflow. Extracting the themes, the reviewers achieved 79% Kappa score, which demonstrates a substantial agreement between the two reviewers.

1) *Topic One: Digital certificate and configuration problems.* The manual analysis for the first topic depicts that developers discussed two main areas, namely certificate/OpenSSL (63%) and SSH (37%). For instance, the discussions were related to OpenSSL configuration, signing and verifying a signature, and generating PEM files using OpenSSL. There were also questions concerning how to generate self-signed certificates, access a certificate store, create a Certificate Signing Request (CSR), establish https and secure connections, and configure certificate-based authentication in ASP.NET. In the SSH-related questions, the majority of the users had difficulty setting SSH with no password, checking the right permission for SSH keys, using SSH programmatically, and connecting to SSH servers of other platforms (*e.g.*, Amazon).

2) *Topic Two: Programming issues.* As for the programming issues topic, we observed that the three most frequently discussed programming languages were Java (*i.e.*, 44), C/C++ (*i.e.*, 31), and C# (*i.e.*, 19). In 31% of the posts developers discussed issues related to the AES algorithm such as different encryption modes (*e.g.*, CBC and ECB) and key sizes (*e.g.*, 128, 192, and 256-bit). In addition to symmetric encryption, 47% of the posts were related to working with asymmetric encryption (*i.e.*, RSA). The challenges were mostly concerned with different padding modes (*e.g.*, OAEP), how to calculate or understand the raw modulus and exponent numbers, and how to generate and work with different key file encodings in RSA (*e.g.*, DER-encoded format, PEM, or XML). Moreover, another evident problem was dealing with different RSA key formats, *i.e.*, Public Key Cryptography Standards (PKCS). The

⁵<https://github.com/bmabey/pyLDAvis>

⁶http://185.94.98.132/~crypto/paper_data/lda.html

⁷http://185.94.98.132/~crypto/paper_data/tags-topics.csv

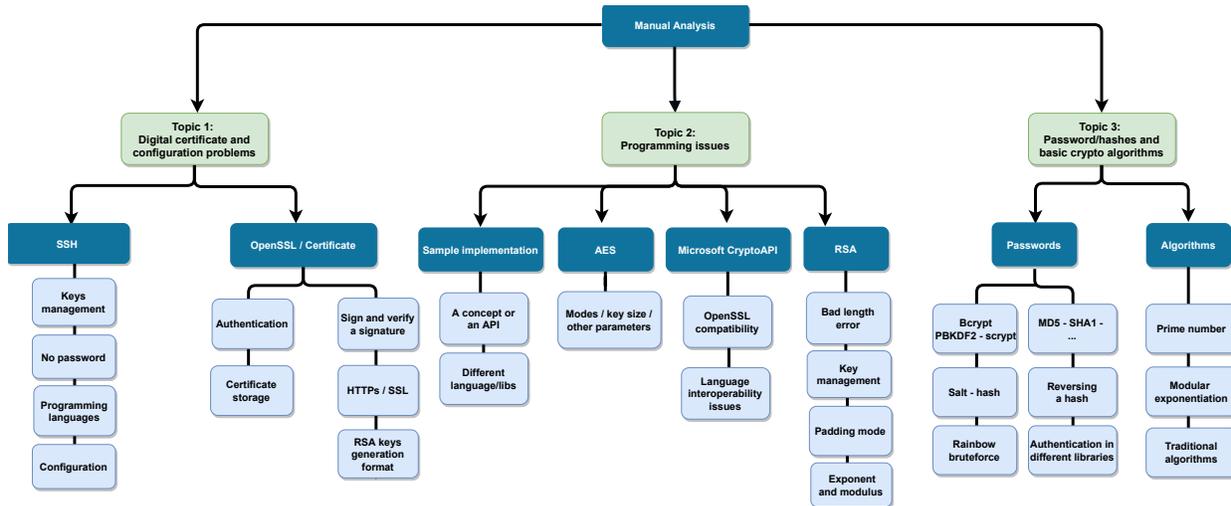


Fig. 1. The results of manual analysis for the three topics

users commonly asked how to convert PKCS#8 to PKCS#1 or other standards, and how to programmatically generate or use different key standards in various crypto libraries (*e.g.*, Bouncy Castle). There were users who had problems with illegal block size errors, often misunderstanding the suitable usage of RSA, *e.g.*, encrypting a long text. Nevertheless, the discussions were resolved by proper responses that suggested incorporating AES and RSA into the encryption/decryption scenario. Another type of question was about the issues in Microsoft CryptoAPI (12%). Developers reported issues on working with OpenSSL or using RSA keys from other sources, *e.g.*, importing keys from OpenSSL into Crypto API, converting RSA keys to be used by Bouncy Castle, verifying an OpenSSL DSA signature using CryptoAPI, having extra fields in generated keys by PHP OpenSSL, and signing a message with pyOpenSSL in Python and verifying it with CryptoAPI. Moreover, there were questions (10%) associated with how to either implement a scenario, *e.g.*, encryption of a string with RSA public key with Swift on iOS, or deal with problems while working with more than one crypto library or programming language, *e.g.*, encryption of a string with RSA in JavaScript and decryption in Java, or decryption of a string in Java which is already encrypted using AES-256 in iOS.

3) **Topic Three: Password/hashes and basic crypto algorithms.** Our findings for the password/hash topic suggest that users primarily discussed problems associated with either passwords (86%) or basic crypto algorithms (14%). Different facets of producing secured passwords were the topic of most discussions. First and foremost, users were uncertain which hashing algorithms (*e.g.*, MD5, SHA-1) can provide a higher level of reliability and how password length contributes to the strength of the resulting hash. Users lacked the required knowledge as to what salt is and how salt can maximize the security of a hash. In addition to pointing out the pros and cons of static salt vs random salt, respondents encouraged users to use salted passwords in order to render the brute-

force or the rainbow table attack prohibitively expensive. Developers were doubtful about which crypto functions, *i.e.*, `bcrypt()`, `PBKDF2()`, or `Scrypt()`, are more secure and faster, and what key differences distinguish the three functions from other hashing algorithms, *e.g.*, MD5, SHA-256. As regards the basic crypto algorithms, users contributed to responses concerning how to produce or find prime numbers, how to use the `BigInteger` class for RSA modular exponentiation, how to produce unique URL safe hash or IDs, and how to solve a Caesar Cipher or substitution ciphers. Lastly, a few users discussed how to program an authentication module in web programming frameworks such as Laravel, or CakePHP.

4) **Topic difficulty and popularity:** We checked the popularity and difficulty level of each topic so as to determine which questions attracted more attention or received acceptable answers with a longer time span, which the same approach was used in the previous study [7]. We used four factors to measure the popularity of a topic, namely the average number of views of documents, the average number of comments, the average number of favorites, and the average score of documents. The four factors can be found in the CSV files,⁸ namely `CommentCount`, `FavouriteCount`, `Score`, and `ViewCount`. We considered the average number of `ViewCount` as the foremost factor to judge the popularity of a topic, the question's score and the number of `FavouriteCount` as the second most important factors, and the average number of comments as the last factor. To find the most difficult topic, we used two factors, namely the average time it takes for a document to obtain an accepted answer, and the ratio of the average number of answers in documents to the average number of the views. We avoided recently posted questions from affecting the analysis by only including those that are older than six months.

We infer that questions related to the usage of digital certificates, and configuration problems are the most popular (highest average `ViewCount` and `FavouriteCount`), and questions

⁸http://185.94.98.132/~crypto/paper_data/

related to hashing and passwords are also viewed as popular based on the other two factors (*i.e.*, average CommentCount and Score). From the difficulty standpoint, we notice that the programming issues topic is the most difficult topic as it had a greater average response time, and its proportion of average answers to average views is the lowest.

5) *Summary*: The challenges in each theme were studied in detail to demonstrate how developers struggle to use or comprehend various areas of cryptography. According to our findings, we believe that there are two foremost reasons with which developers mainly encounter problems in cryptography. The first leading cause is a distinct lack of knowledge to discern *why* or *what* they need to use to accomplish a crypto task. We observed ample evidence where developers lacked the confidence to choose the best algorithm or parameter, for instance, the right and safest padding option in AES. Consequently, developers may use boilerplate code snippets from the provided answers, in spite of the answers' reliability and security. In the second factor, although the fundamental concepts are the same, the implementation approach of a crypto concept in various crypto libraries is influential to developer performance. Compelling evidence in findings urges that working with more than a crypto library due to using various architectures or platforms in a project creates confusion for developers regarding *how* a particular problem can be resolved. They commonly have trouble in creating keys with one library and import them into another library or verifying a signature in a different crypto library. Furthermore, adequate explanations and the existence of useful examples in documentations can alleviate the difficulty of using cryptography.

V. THREATS TO VALIDITY

In this study, we concentrate on one major platform where developers discuss crypto topics. This may not be sufficient as there are many other platforms, such as crypto Stack Exchange, which can provide more data to analyze. We measured topic difficulty and popularity based on metrics used in the previous study. Nevertheless, these observations may not be sufficient to determine what type of crypto questions are more challenging than others. Users may not always feel responsible for selecting a reasonable answer as an acceptable answer. Therefore, not having an accepted answer does not necessarily determine if the question is challenging for others.

VI. CONCLUSION

We conducted a large-scale study on crypto issues discussed on Stack Overflow to find out what crypto challenges users commonly face in various areas of cryptography. Findings suggest that developers still have a distinct lack of knowledge of fundamental concepts, such as OpenSSL, asymmetric and password hashing, and the complexity of crypto libraries weakened developer performance to correctly realize a crypto scenario. We call for dedicated studies to investigate the usability of crypto APIs. We are conducting a survey with users who actively helped the Stack Overflow community in this domain to understand the potential remedies.

Acknowledgments

We gratefully acknowledge the financial support of the Swiss National Science Foundation for the project "Agile Software Assistance" (SNSF project No. 200020-181973, Feb. 1, 2019 - April 30, 2022).

REFERENCES

- [1] S. Nadi, S. Krüger, M. Mezini, and E. Bodden, "Jumping through hoops: Why do java developers struggle with cryptography apis?" in *Proceedings of the 38th International Conference on Software Engineering*.
- [2] M. Hazhirpasand, M. Ghafari, and O. Nierstrasz, "Java cryptography uses in the wild," in *Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 2020, pp. 1–6.
- [3] Y. Tymchuk, M. Ghafari, and O. Nierstrasz, "Jit feedback - what experienced developers like about static analysis," in *2018 IEEE/ACM 26th International Conference on Program Comprehension (ICPC)*, 2018, pp. 64–6409.
- [4] C. Corrodi, T. Spring, M. Ghafari, and O. Nierstrasz, "Idea: Benchmarking android data leak detection tools," in *Engineering Secure Software and Systems*, M. Payer, A. Rashid, and J. M. Such, Eds. Cham: Springer International Publishing, 2018, pp. 116–123.
- [5] S. Kafader and M. Ghafari, "Fluentcrypto: Cryptography in easy mode," in *37th International Conference on Software Maintenance and Evolution (ICSME)*, 2021.
- [6] S. E. Jahan, M. Rahman, A. Iqbal, and T. Sabrina, "An exploratory analysis of security on data transmission on relevant software engineering discussion sites," in *2017 4th International Conference on Networking, Systems and Security (NSysS)*. IEEE, 2017.
- [7] X.-L. Yang, D. Lo, X. Xia, Z.-Y. Wan, and J.-L. Sun, "What security questions do developers ask? a large-scale study of stack overflow posts," *Journal of Computer Science and Technology*, vol. 31, no. 5, 2016.
- [8] N. Meng, S. Nagy, D. Yao, W. Zhuang, and G. A. Argoty, "Secure coding practices in java: Challenges and vulnerabilities," in *Proceedings of the 40th International Conference on Software Engineering*, 2018.
- [9] F. Fischer, H. Xiao, C.-Y. Kao, Y. Stachelscheid, B. Johnson, D. Razar, P. Fawkesley, N. Buckley, K. Böttinger, P. Muntean *et al.*, "Stack overflow considered helpful! deep learning security nudges towards stronger cryptography," in *28th {USENIX} Security Symposium*, 2019.
- [10] F. Fischer, K. Böttinger, H. Xiao, C. Stransky, Y. Acar, M. Backes, and S. Fahl, "Stack overflow considered harmful? the impact of copy&paste on android application security," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017.
- [11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, 2003.
- [12] A. A. Bangash, H. Sahar, S. Chowdhury, A. W. Wong, A. Hindle, and K. Ali, "What do developers know about machine learning: a study of ml discussions on stackoverflow," in *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*. IEEE, 2019.
- [13] C. Rosen and E. Shihab, "What are mobile developers asking about? a large scale study using stack overflow," *Empirical Software Engineering*, vol. 21, no. 3, 2016.
- [14] H. Osman, M. Ghafari, and O. Nierstrasz, "Hyperparameter optimization to improve bug prediction accuracy," in *2017 IEEE Workshop on Machine Learning Techniques for Software Quality Evaluation*, 2017.
- [15] W. Zhao, J. J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, and W. Zou, "A heuristic approach to determine an appropriate number of topics in topic modeling," in *BMC bioinformatics*, vol. 16, no. 13, 2015.
- [16] G. Luo, "A review of automatic selection methods for machine learning algorithms and hyper-parameter values," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 5, no. 1, 2016.
- [17] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, 2004.
- [18] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in neural information processing systems*, 2009.
- [19] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, vol. 3, no. 2, 2006.
- [20] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, 1960.