Automated and Personalized Privacy Policy Extraction under GDPR Consideration

Cheng Chang, Huaxin Li, Yichi Zhang, Suguo Du, Hui Cao, and Haojin Zhu

Shanghai Jiao Tong University, Shanghai, China

Abstract. Along with the popularity of mobile devices, people share a growing amount of personal data to a variety of mobile applications for personalized services. In most cases, users can learn their data usage from the privacy policy along with the application. However, current privacy policies are always too long and obscure to provide readability and comprehensibility to users. To address this issue, we propose an automated privacy policy extraction system considering users' personal privacy concerns under different contexts. The system is implemented on Android smartphones and evaluated feedbacks from a group of users ($n = 96$) as a field study. Experiments are conducted on both our dataset, which is the first user privacy concern profile dataset to the best of our knowledge, and a public dataset containing 115 privacy policies with 23K data practices. We achieve 0.94 precision for privacy category classification and 0.81 accuracy for policy segment extraction, which attests to the significance of our work as a direction towards meeting the transparency requirement of the General Data Protection Regulation (GDPR).

Keywords: privacy policy extraction · GDPR · mobile application privacy

1 Introduction

In recent years, we have witnessed a huge growth in emerging mobile techniques such as 5G communication and Internet of Things (IoT). These techniques have empowered the functionalities of mobile devices, and thus service providers (e.g., device OEMs, mobile apps developers) are able to provide more ubiquitous, seamless, and personalized applications to users. However, many of these *mobile applications* (e.g., user-profile-based recommendation, real-time location services, financial apps) widely rely on personal or private information, which brought growing privacy attentions. The situation can be more serious when personal data are shared with the third party for the advertising purpose [12,14,21]. To secure users' privacy, one of the critical steps is to let users be aware of potential privacy risks that can be brought by using a mobile app.

Until now, privacy policies of mobile apps (e.g., Android apps, web apps) are still the primary channels through which users are able to know their personal data usage. They describe the detailed usage of data collected by apps from a range of aspects, including what and how the data is collected, data security,

data retention, user access and control, whether the data is shared with the third party and so on. Ideally, it should be totally decided by users to accept or refuse the policy. In practice, unfortunately, most users just omit to read the privacy policies, as they are always too long and complex [4,15]. Previous works demonstrated the current design of the privacy policy breaches its original intention. The problem is more pressing after the General Data Protection Regulation (GDPR) went into effect in 2018, as it presents transparency requirement in article 12 that *the controller shall take appropriate measures to provide any information* related to processing the data *in a concise, transparent, intelligible and easily accessible form, using clear and plain language* [19].

Several frameworks were proposed to help fill the gap between the current privacy policy design and the transparency requirement of the GDPR, and most of them focus on the fine-grained policy segment classification to reorganize the policy in a more noticeable security-centric presentation [5,20]. However, users still need to have enough background knowledge about pre-defined terminology and information structure to search and scrutinize their concerned privacy issue. These works are also GDPR-agnostic, causing they lack legal warning to both app providers and users.

In this work, we propose a practical solution that learns users' privacy concerns and automatically extracts corresponding descriptions in privacy policies when users use different kinds of mobile apps. To achieve the goals, the first challenge is to build the user privacy concern profile, since no previous work has provided related datasets. We collect our privacy concern dataset through crowdsourcing and interviewing. Then we aggregate the individual privacy profiles by hierarchical clustering and design a matching mechanism to assign one of the profile clusters to a new user. The second challenge is to automatically analyze privacy policies on a large scale and extract policy segment accurately. To accomplish this task, we deploy a deep Convolutional Neural Network (CNN) followed by a random forest as the core of the policy extraction module. We train the model on OPP-115 dataset [20] containing 115 privacy policies with 23K fine-grained manual annotations, and achieve considerable performance.

Our main contributions are summarized as follows.

- We design and implement a system, which is composed of user privacy profile generation and privacy policy extraction modules. This system automatically provide a user with the descriptions in the app's privacy policy which she most cares about, as well as related GDPR items in order to help her make decisions with enough privacy awareness.
- We build the first dataset depicting the user privacy concern profiles. The dataset is constructed with 252 participants. We further design a matching profile assignment method, which succeeds to fast provide users with according profile identifications in the field study.
- We demonstrate our deep learning based privacy policy extraction model achieving 0.88 F1-score on the OPP-115 datatset [20], which outperforms the state-of-the-art privacy policy analysis system [5].

– We conduct the field study with 96 participants to comprehensively assess our system in practice, where the system achieves 0.81 accuracy on privacy policy presentation.

2 Related Work

Privacy Policy Analysis. Prior work has explored the methods on improving the readability and comprehensibility of privacy policies. Zimmeck *et al.* [23] proposed to classify privacy policies by machine learning and built an automatic analyzing architecture. Holtz *et al.* [6] designed simple and obvious icons to represent the contents of privacy policies. However, their taxonomy is somewhat coarse and classification accuracy is relatively low. To this end, Wilson *et al.* [20] created the first public privacy policy corpus and employed three experts to manually label the policies with fine-grained annotations. After that, researchers further implemented online tools to support querying on privacy policies in practice. Oltramari *et al.* [16] designed a semantic framework to visualize the structure of privacy policies. Sathyendra *et al.* [17] proposed an approach towards automatically detecting the provision of choices in privacy policies by NLP techniques. As we know until now, Harkous *et al.* [5] presented the most comprehensive system to enable scalable and multi-dimensional privacy policy analysis. But they all rely on active querying and searching by users and require users to have related background knowledge. Our work is the first to automatically predict mobile users' privacy concerns and then extract the target parts of privacy policy with GDPR consideration.
GDPR Influence. Since the GDPR, a regulation setting a high standard for personal data processing, went into effect in 2018, researchers started to analyze the influence of the GDPR on the current circumstance of privacy policies. Linden *et al.* [10] discovered that, due to the transparency requirement by the GDPR, privacy policies are shown in a more organized structure and have greater number of words. Degeling *et al.* [3] accessed 6357 websites in total and found a 4.9% increase in the number of websites owning a privacy policy. Tesfay *et al.* [18] made an attempt to assess a privacy policy with the criteria of risk level on violating the GDPR. Their work reveals the positive influence of the GDPR. Therefore, we propose to extract related GDPR items to improve the user's privacy awareness.

3 Framework Overview

Fig. 1 shows an overview of our system. It comprises two modules: the privacy concern profile generation module and the privacy policy extraction module. The former module is responsible for generating personalized privacy concern profile when a user first signs in the system. We emphasize *user's privacy concerns*, defined as ten categories of security-centric data usage in privacy policies (listed in Table 1). For instance, if a user declares that she worries about whether her personal information is misused by online social apps, *First Party Collection* is

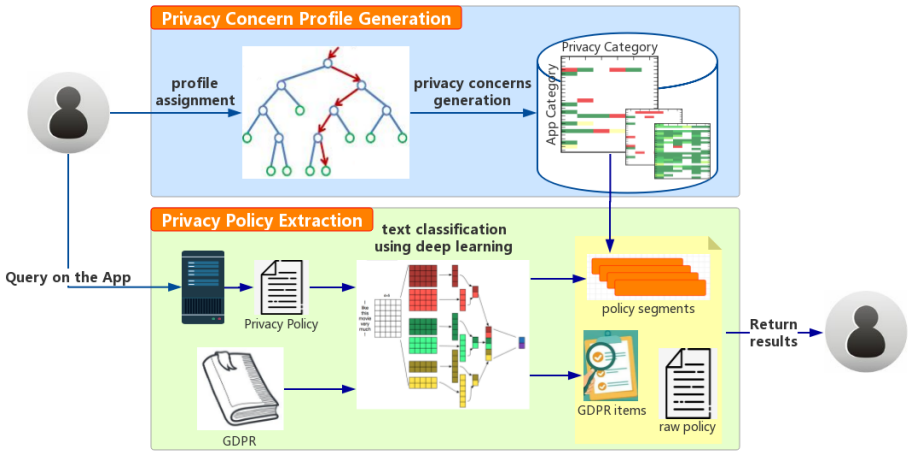

Fig. 1. The high-level overview of the system.

then considered as one of her privacy concerns. The latter module utilizes the strength of deep learning to extract the text segments of the privacy policy and GDPR items, according to the profile generated before. A demo[1] is released to introduce the functionality of our system.

Privacy Concern Profile Generation. This module generates the personalized privacy concern profile for an user in the interactive question-answering form. The user will be asked at most 5 questions about her privacy concerns while using different kinds of apps. The latter question is dynamically elicited by the former answer. According to the answers, a learned profile which is clustered from the profile database is assigned to the user. Then the concern profile is used for the personalized privacy policy extraction.

Privacy Policy Extraction. This module extracts the target segments in the app's privacy policy queried by the user. In this module, a deep learning pipeline, which is comprised of Convolutional Neural Network (CNN) and random forest model, takes as input the segments of the privacy policy scraped from the server and the GDPR. Then the model labels each segment with a set of category-attribute values describing its privacy-related data practice. According to the user profile generated in the previous module, the GDPR-aware query result is the combination of the descriptions in the privacy policy which the user most cares about and the related regulation items in the GDPR. We illustrate the output of our system using an actual example in Appendix A.

4 Privacy Concern Profile Generation

To build the privacy concern profile dataset, we recruited 252 participants on the crowdsourcing platform, Amazon Mechanical Turk (MTurk) [8], to complete

[1] `https://youtu.be/-0x-HQRnYwQ`

the questionnaire[2] about personal privacy concerns. We cluster the profiles and design a mechanism to assign one of the profiles to a new user.

4.1 Dataset Collection

Questionnaire Design. In order to know a user's personal privacy concerns under different contexts, we ask the participants to list which aspects of privacy they most care about in order while using different kinds of mobile apps. We use app categories from the Google Play and privacy taxonomy from the public dataset OPP-115 [20]. Table 1 describes the adopted privacy categories.

Table 1. the privacy category and description in OPP-115.

Privacy Category	Description
First Party Collection	how the app collects user data.
Third Party Sharing	how the user data is shared with third parties.
User Choice/Control	choices and control options the app grants to users.
User Access and Edit	how users can access or edit their data.
Data Retention	how long the app stores the user data.
Data Security	how the app protects the user data.
Policy Change	how the app informs users about privacy policy changes.
Do Not Track	how DNT signals for online tracking is honored.
Specific Audiences	particular policies to some specific groups of users.
Other	introductory information.

Participant Recruitment. We also collect users' demographic features and IUIPC score [13] to make the bias of the dataset as low as possible.

we recruited 252 MTurk workers with *Master* qualification and approval rate more than 85%. We paid each participant $2 for the work. On average, the survey lasts about 25 minutes, which means the participants answered the questions with enough consideration as our expectation. These participants come from different areas (North America 70.2%, Asia 16.7%, South America 10.3%, Europe 2.4%, and Africa 0.4%), have different genders (62.7% male and 37.3% female), own different education degrees (Bachelor's degree 61.1%, graduate degree 13.1%, associate degree or lower 25.8%), and are at different ages (23-30 years old 45.6%, 30-40 years old 31.0%, beyond 40 years old 21.0%, under 22 years old).

IUIPC is a 10-item scale measure of the user's privacy awareness. In each item, the awareness is scaled from 1 to 7. Its effectiveness is demonstrated by some statistic criteria and it is widely used in previous works on privacy research [1,11]. According to [13], people who have stronger privacy awareness should have higher scores in IUIPC. These recruited participants have an average score of 5.79 and 67.2% of the participants get a score >6 (the maximum score is 7),

[2] `https://www.wjx.top/jq/33235531.aspx`

which shows these participants have strong privacy awareness and thus probably care about the privacy policies.

4.2 Profile Generation

From the dataset, the detailed user privacy profile is constructed in the matrix form, where each row corresponds to a category of apps and the columns represent different privacy category listed in Table 1. If a participant reports that she most cares about privacy j while using apps of category i, the corresponding value at index *(i,j)* in the matrix is set to 1. Otherwise, the value is set to 0.
Profile Clustering. We apply the technique of hierarchical clustering on the profile matrices. The reason for choosing the hierarchical clustering is that it is non-parametric and able to provide visualized explanations.

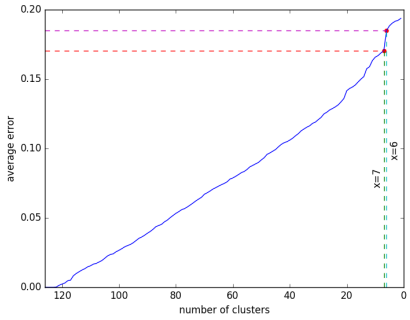

Fig. 2. The average error changes in clustering process.

Fig. 2 shows the average error curve while clustering using the bottom-up strategy. The error is calculated by averaging the difference between each profile matrix P and its corresponding cluster C per position, i.e.,

$$cluster\_error = \sum_{\forall(i,j)} |P(i,j) - C(i,j)| \tag{1}$$

From Fig. 2, we discover that the average error rapidly increases when the remained seven clusters are clustered into six (0.170 to 0.185), which means the information loss is too high to merge the two clusters. Therefore, we choose these seven clusters as typical privacy concern profiles, as it meets both lower information loss and higher profile representativeness.
Profile Assignment. The consequence of the profile clustering is displayed in Fig. 3. Totally, we obtain seven profiles revealing the diversity of privacy concerns. The majority of participants are clustered into Profile 1 and Profile 2, correspondingly occupies 42.9% and 41.3% of the whole. Profile 1 represents participants who most care about data security for most apps. Profile 2 contains

the ones who show solicitude for the data collected by the first and the third parties. Profile 6 emphasizes the rights of user control. Profile 7 focuses on data retention and data security for apps related to financial and location information. Profile 3, 4, 5 seem not to mind their privacy, except for some special cases. Profile 5 is like the privacy careless counterpart of Profile 1.

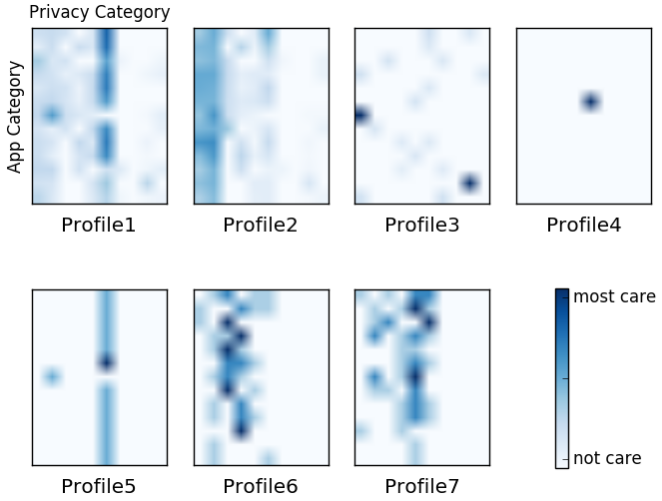

Fig. 3. The clustering results of the privacy profiles.

To assign one of these profiles to a new user, we craft a dynamic interactive question-answering mechanism, where the user is asked at most 5 yes/no questions capturing the discriminative features of profiles, to be eventually assigned to a certain profile. Depending on the user's privacy preferences, the former answer decides the latter question, so different users may answer different set of questions to better personalize their profiles.

5 Privacy Policy Extraction

To extract the concerned information, we deploy a deep learning model as the classifier to label the text segments of the privacy policy. According to the user profile generated as described in section 4, the expected segments are returned to the user, together with related GDPR items to reinforce the privacy awareness.

5.1 Dataset Description

We leverage the public dataset OPP-115 [20] to train our deep learning classifier. The dataset contains 115 privacy policies with 23K fine-grained annotations manually labeled by three experts. The annotation scheme is at two levels. Level

1 annotates each paragraph-sized segment with one or more of ten privacy categories in Table 1. Level 2 is a group of <attribute-value> pairs concretely illustrating the privacy data practice. For instance, if a segment is annotated as *First Party Collection* in level 1, it must contain 3 mandatory attribute annotations: *Collection Action*, *Information Type* and *Purpose* in level 2. Each attribute has a value coming from its own defined value set. For example, the value set of *Purpose* attribute is: *basic services*, *advertising*, *research*, etc. In total, there are 18 distinct mandatory attributes across all categories, and the size of value sets of each attribute is between 4 and 16.

5.2 Classifier Model

The classifiers classify privacy categories and predict values of attributes for policy text segments. The privacy categories are used to match the privacy concern profiles of users so that we can render the policy segments they most care about.
Model Hierarchy. Due to the two-level hierarchical nature of the data label, we train the classifiers at both levels inspired by the previous work [20]. At the first level, there is one unique classifier predicting the probability of the privacy category $p(c_i|s), c_i \in C$, where s is the input segment and C is the set of all categories in Table 1. At the second level, there are 18 classifiers corresponding to 18 attributes. Each classifier predicts the value describing the attribute $p(v_j|s), v_j \in V(a)$, where $V(a)$ is the set of possible values for a single attribute a.

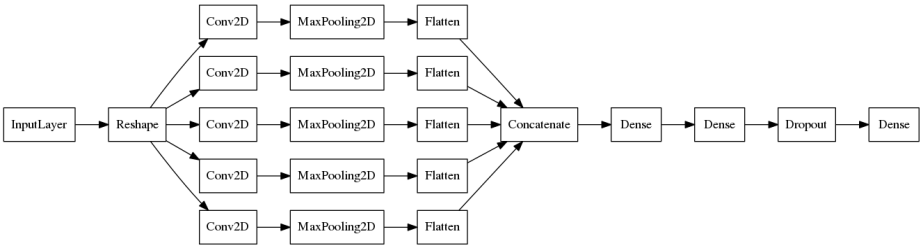

Fig. 4. The CNN structure contains five kinds of filter sizes (from 2 to 6). Each filter size contains a convolutional layer followed by a max-pooling layer. Two fully-connected layers are added after the max-pooling layers. The active function is *sigmoid* for the output layer and *Relu* for convolutional layers and fully-connected layers. A dropout layer is applied to avoid over-fitting.

Classifier Construction. Our classifier contains a CNN followed by a random forest. Fig. 4 illustrates the structure of the CNN. Previous work has demonstrated that CNN is appropriate for tasks of text classification [7,9,22]. The output of the CNN, the probability distribution of categories or attribute values, is then fed to a random forest. We constrain the depth of trees no more than 4 to avoid over-fitting.

The model is implemented in Python. A segment is tokenized and embed to a list of word vectors using the *fastText* library [2]. The CNN is constructed by the *Keras*[3] API and the random forest is built using the *sklearn*[4] toolkit.

Model Performance. We randomly select 85 privacy policies in the dataset for training and the remaining 30 policies for testing. The hyper-parameters are obtained by grid-search and 5-fold cross-validation. The CNN converges after 300 epochs.

From Table 2, we can see that on average, our model achieves 0.94 precision, 0.83 recall, and 0.88 F1-score, which are higher than *Polisis*, the state-of-the-art privacy policy analysis system [5]. We also compare our model with approaches of SVM and hidden Markov model (HMM) adopted in OPP-115 as baselines. The results show our model outperforms them with 0.15 to 0.20 degree of metrics.

Table 2. Test results of the category classifier. Hypermeters: word-vector dimension: 300, number of filters for each convolutional layer: 100, intermediate fully-connected layer size: 100, 20, dropout rate: 0.5. batch size: 32.

Privacy Category	Precision	Recall	F1-score
First Party Collection	0.94	0.88	0.91
Third Party Sharing	0.92	0.85	0.89
User Choice/Control	0.97	0.71	0.82
User Access and Edit	0.99	0.80	0.89
Data Retention	0.99	0.61	0.76
Data Security	0.94	0.74	0.83
Policy Change	0.88	0.68	0.77
Do Not Track	0.99	0.80	0.89
Specific Audiences	0.99	0.87	0.93
Other	0.92	0.86	0.89
Average	0.94	0.83	0.88
Polisis [5]	0.87	0.83	0.84
SVM	0.66	0.66	0.66
HMM	0.60	0.59	0.60

6 Field Study

We conduct the field study to validate the effectiveness of our system in practice. In this study, we implement our app on Android platform, and deploy the server for privacy policy scraping and text segment extraction. The participants ($n = 96$) are recruited from MTurk. They are asked to assess if our prediction hits the center of their privacy concerns.

Study Procedure. The main goal of this field study is to demonstrate the accuracy of our methods and the feasibility of the system. As described in section 4.2,

[3] `https://keras.io`

[4] `https://scikit-learn.org/`

A participant must pass the dynamic question-answering process to be assigned to one of the seven learned privacy profiles. The questions have two forms: (1) privacy category only, for example, "Do you most care about the *Data Security* contents?" (2) <app category, privacy category> pair, for example, "Do you most care about the *First Party Collection* contents when you use social apps?".

After the assignment, the app connects to the server and transmits the profile index. For each app category, we randomly choose one app from top-50 hots on the Google Play. Its privacy policy is automatically downloaded and divided into segments on the server. At the time, the two-level classification model launches to label all the segments and then extract the concerned ones according to the profile. Subsequently, the raw segments are transmitted to the app, together with related GDPR items. The participants are responsible for judging if the prediction hits their privacy concerns. Finally, the feedback is reported to the server. At the end of the study, the participants are requested to complete the same survey as raised in section 4.1. Each participant is paid $3 after the task.

Study Results. In total, we collect 1910 segment reviews across all the app categories. Fig. 4 shows the concrete feedback aggregated by the privacy category. Overall, there are 1548 correctly predicted segments, which means the accuracy is around 0.81. Meanwhile, our system performs extremely well (0.85 accuracy) on the most focused three privacy category (occupying 87% segments).

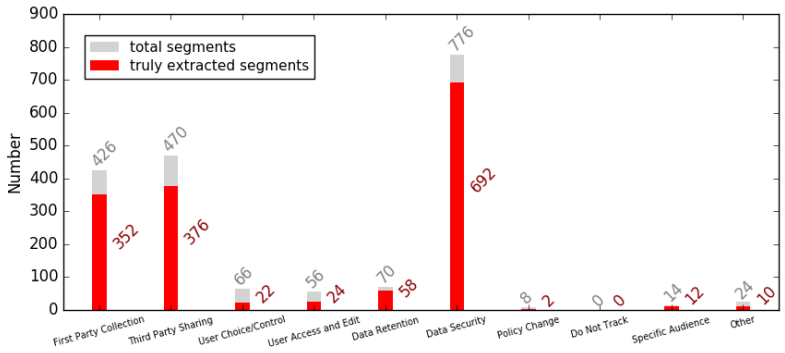

Fig. 5. The concrete feedback results aggregated by the privacy category.

7 Conclusion

In this paper, we create the first system to automatically predict and extract app's privacy policies with personalized privacy concerns. The system is composed of two modules: users' privacy concern profile generation and privacy policy extraction. To generate the profile depicting users' privacy concerns under

different contexts, we construct the first dataset and design a matching mechanism to fast assign a learned profile to a user. Then, we deploy a deep learning based NLP model to recognize and elicit the target descriptions in app's privacy policy and related items in the GDPR. The real-world field study demonstrates the effectiveness of our system, where we accurately provide the users with the contents under concerns at 0.81 accuracy.

References

1. Angst, C.M., Agarwal, R.: Adoption of electronic health records in the presence of privacy concerns: The elaboration likelihood model and individual persuasion. MIS quarterly **33**(2), 339–370 (2009)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
3. Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., Holz, T.: We value your privacy... now take some cookies: Measuring the gdpr's impact on web privacy. arXiv preprint arXiv:1808.05096 (2018)
4. Gluck, J., Schaub, F., Friedman, A., Habib, H., Sadeh, N., Cranor, L.F., Agarwal, Y.: How short is too short? implications of length and framing on the effectiveness of privacy notices. In: 12th Symposium on Usable Privacy and Security (SOUPS). pp. 321–340 (2016)
5. Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K.G., Aberer, K.: Polisis: Automated analysis and presentation of privacy policies using deep learning. In: 27th {USENIX} Security Symposium ({USENIX} Security 18). pp. 531–548 (2018)
6. Holtz, L.E., Zwingelberg, H., Hansen, M.: Privacy policy icons. In: Privacy and Identity Management for Life, pp. 279–285. Springer (2011)
7. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
8. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with mechanical turk. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp. 453–456. ACM (2008)
9. Li, P., Zhao, F., Li, Y., Zhu, Z.: Law text classification using semi-supervised convolutional neural networks. In: 2018 Chinese Control And Decision Conference (CCDC). pp. 309–313. IEEE (2018)
10. Linden, T., Harkous, H., Fawaz, K.: The privacy policy landscape after the gdpr. arXiv preprint arXiv:1809.08396 (2018)
11. Liu, B., Andersen, M.S., Schaub, F., Almuhimedi, H., Zhang, S.A., Sadeh, N., Agarwal, Y., Acquisti, A.: Follow my recommendations: A personalized privacy assistant for mobile app permissions. In: Symposium on Usable Privacy and Security (2016)
12. Ma, Z., Wang, H., Guo, Y., Chen, X.: Libradar: fast and accurate detection of third-party libraries in android apps. In: Proceedings of the 38th international conference on software engineering companion. pp. 653–656. ACM (2016)
13. Malhotra, N.K., Kim, S.S., Agarwal, J.: Internet users' information privacy concerns (iuipc): The construct, the scale, and a causal model. Information systems research **15**(4), 336–355 (2004)
14. Mayer, J.R., Mitchell, J.C.: Third-party web tracking: Policy and technology. In: Security and Privacy (SP), 2012 IEEE Symposium on. pp. 413–427. IEEE (2012)

15. McDonald, A.M., Cranor, L.F.: The cost of reading privacy policies. ISJLP **4**, 543 (2008)
16. Oltramari, A., Piraviperumal, D., Schaub, F., Wilson, S., Cherivirala, S., Norton, T.B., Russell, N.C., Story, P., Reidenberg, J., Sadeh, N.: Privonto: A semantic framework for the analysis of privacy policies. Semantic Web pp. 1–19 (2017)
17. Sathyendra, K.M., Wilson, S., Schaub, F., Zimmeck, S., Sadeh, N.: Identifying the provision of choices in privacy policy text. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2774–2779 (2017)
18. Tesfay, W.B., Hofmann, P., Nakamura, T., Kiyomoto, S., Serna, J.: I read but don't agree: Privacy policy benchmarking using machine learning and the eu gdpr. In: Companion of the The Web Conference 2018 on The Web Conference 2018. pp. 163–166. International World Wide Web Conferences Steering Committee (2018)
19. the General Data Protection Regulation. `https://gdpr-info.eu/`
20. Wilson, S., Schaub, F., Dara, A.A., Liu, F., Cherivirala, S., Leon, P.G., Andersen, M.S., Zimmeck, S., Sathyendra, K.M., Russell, N.C., et al.: The creation and analysis of a website privacy policy corpus. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1330–1340 (2016)
21. Zhang, L., Cai, Z., Wang, X.: Fakemask: A novel privacy preserving approach for smartphones. IEEE Transactions on Network and Service Management **13**(2), 335–348 (2016)
22. Zhang, Z., Zou, Y., Gan, C.: Textual sentiment analysis via three different attention convolutional neural networks and cross-modality consistent regression. Neurocomputing **275**, 1407–1415 (2018)
23. Zimmeck, S., Bellovin, S.M.: Privee: an architecture for automatically analyzing web privacy policies. In: Proceedings of the 23rd USENIX conference on Security Symposium. pp. 1–16. USENIX Association (2014)

A An Example of the System Output

If a user queries the privacy policy of the *WeChat* app, and her privacy concern profile shows solicitude for whether her personal data is secure. Then the system will return a text segment in the *WeChat* privacy policy:

We use a variety of security technologies and procedures for the purpose of preventing loss, misuse, unauthorised access, or disclosure of Information – for example... But *no data security measures can guarantee 100% security at all times. We do not warrant or guarantee the security of WeChat or any information you provide to us through WeChat.*

And the related regulation items in GDPR is also presented to the user:

...the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, including inter alia as appropriate:...