Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

SensitivAlert: Image Sensitivity Prediction in Online Social Networks Using Transformer-based Deep Learning Models

Lindrit Kqiku, Delphine Reinhardt

University of Göttingen, Institute of Computer Science, Computer Security and Privacy, Germany
University of Göttingen, Campus Institute Data Science (CIDAS), Germany
kqiku@cs.uni-goettingen.de, reinhardt@cs.uni-goettingen.de

Abstract

Billions images are shared daily on social networks. When shared with an inappropriate audience, user-generated images can, however, compromise users' privacy and may have severe consequences, such as dismissals. To address this issue, different solutions were proposed, ranging from graphical user interfaces to *Deep Learning* (DL) models to alert users based on image sensitivity prediction. Although these models show promising results, they are evaluated on datasets relying on small participants' samples. To address this limitation, we first introduce SensitivAlert, a dataset that re-annotates the previously annotated images from two existing datasets, but using a German-speaking cohort of 907 participants. We then leverage it to classify images according to two sensitivity classes—private or public—using recent transformer-based DL models. In our evaluation, we first consider consensus-based generic models using our dataset as benchmark based on image content itself and its associated user tags. Moreover, we show that our fine-tuned models trained on our dataset better reflect users' image privacy conceptions. We finally focus on individual user's privacy estimation by investigating three approaches: (1) a generic approach based on participants' consensus for fine-tuning, (2) a user-wise approach based on user's privacy preferences only, and (3) a hybrid approach that combines individual preferences with consensus-based preferences. Our results finally show that the generic and hybrid approaches outperform the user-wise one for most users, thus ensuring the feasibility of image privacy prediction preferences at the individuals' level.

Introduction

Internet users share daily an unprecedented volume of self-generated content on *Social Network Sites* (SNS). According to (SocialPilot 2023), 36% of Facebook and 72% of Instagram posts are images. Posting these images may threaten the privacy of users, if shared with an inappropriate audience. While SNS may offer interfaces to control their access, it has been shown that users rarely leverage them. The reasons behind it are multiple. For example, these control mechanisms may not be directly accessible by the users (Chen et al. 2019), time-consuming to use (Lipford, Besmer, and Watson 2008), or complex to apply so that they match users' privacy preferences (Alan et al. 2022). Not exercising such control can, however, have severe consequences. For example, the inference of personal information, e.g., from images, can result in cyberstalking (Dressing et al. 2014), doxing (Snyder et al. 2017), or sextorsion (Yates 2017). Instead of relying on a manual control of the audience by the participants, Reinhardt et al. (2015) proposed a privacy assistant that supports users in automatically recognizing sensitive images and alerts them, before being shared on SNS. As basis for its realization, Zerr et al. (2012) introduced *PicAlert* dataset, which is used by several works as a benchmark to investigate image sensitivity prediction. Aside from PicAlert, (Zhao et al. 2022) introduced *PrivacyAlert*. The resulting models are based on a consensus idea of privacy, i.e., on a majority agreement of annotations from several users for a particular image. They assume that such generic models would recognize common patterns of different users' privacy perceptions. However, SNS users often have different and subjective privacy perceptions and concerns (Coopamootoo and Gross 2017). Moreover, PicAlert is based on annotated images by only 81 users recruited from a Computer Science campus, Facebook, and two Russian forums. Besides, PrivacyAlert is based on users recruited from Amazon MTurk and considers only 1,704 private images. Thus, the small users' size may not be representative for SNS users and the small number of private images may lead on building models that do not generalize well for other private images. Additionally, an MTurk cohort is expected to include mainly USA-based participants (75%) and Indians (16%) (Difallah, Filatova, and Ipeirotis 2018). Privacy is, however, a cultural construct (Lunheim and Sindre 1993) as confirmed by inter-cultural differences observed in different domains (EU Commission 2019; Coopamootoo 2020; Murmann et al. 2021; Markos, Milne, and Peltier 2017). Thus, the performances measured in PicAlert and PrivacyAlert may not be representative for other cohorts.

In this paper, we hence investigate different consensus-based and participant-wise sensitivity classification approaches using advanced transfer-learning techniques in a more representative dataset. We fine-tune and evaluate them within our cohort. Moreover, we consider models based not only on image content, but also on user tags and their combinations with deep features. We also evaluate the differences between cohorts by investigating whether our models generalize better when fine-tuned on other cohorts and vice versa.

In summary, we contribute as follows:

- **Our dataset.** We have conducted a user study with 907 German-speaking participants to create a new dataset referred to as *SensitivAlert*. By focusing on a German cohort, we aim to capture in particular the nuances of privacy perceptions within an isolated cohort. Each participant annotated 60 images. Each image was annotated multiple times. From them, we derive a consensus dataset with images annotated with the same label by the majority of their respective annotators.
- **Performance of generic-based models on consensus images.** We fine-tune *BERT Pre-Training of Image Transformers* (BEiT) (Bao et al. 2021), EVA-02 (Fang et al. 2023), and ConvNeXt V2 (Woo et al. 2023) models to classify images according to users' perceived image sensitivity. We combine user tags and deep features generated by pre-trained EVA-02 to fine-tune BERT classification model. We also leverage user tags and deep features jointly to fine-tune ALBEF. The models determine the image sensitivity based on both the user and deep features. We also fine-tune BERT with user tags only resp. deep features only to estimate their individual performance. For example, we reach a 77.48 % f1 on our overall SensitivAlert dataset for our best performing model. Moreover, we compare the performance of our fine-tuned models by (1) training in our cohort and evaluating in each of existing cohorts, and vice versa, against (2) the training and evaluating of those models in our dataset in order to identify potential modelling differences between cohorts. Our results show that fine-tuning BEiT based on our cohort and evaluating on existing ones lead to better results than conversely, with up to 22.41% f1 difference, thus suggesting that our model represents better the privacy preferences of SNS users than others.
- **Participant-wise generic model performance.** We further explore the performance of the generic model at the individual level by fine-tuning and validating BEiT and EVA-02 models on the consensus-based dataset and evaluating the performance for each participant in the evaluation set. BEiT reaches 0.69 f1 average participant-wise, demonstrating its feasibility on individuals.
- **Performance of personalized-based models.** We also evaluate the performance of user-wise and hybrid image sensitivity classification approaches. The former is based on user-specific annotations only. For each participant, we hence fine-tune and validate a separate model in a subset of his/her annotations, and evaluate each model in the remaining annotations. The latter combines user-specific annotations with the consensus ones by fine-tuning BEiT on the consensus dataset and a subset of user-wise annotations from the individuals' annotations set. Thus, we are able to compare the performance of generic, user-wise and hybrid image sensitivity prediction approaches based on images derived from the same cohort. Our results show that the hybrid approach outperforms the user-wise approach for most users.

The rest of the paper is structured as follows: We first discuss related work. We then introduce our data collection methodology, our dataset, and detail our baseline modeling and approaches. We further highlight our results and discuss them, before concluding and discussing future work.

Related Work

We categorize related work as follows: (1) Inferring the sensitivity of different information types (not images) in diverse countries, (2) image sensitivity datasets and image taxonomies based on literature reviews, and (3) *Machine Learning* (ML)-based image sensitivity prediction. The latter thus shares the most similarities with our work.

Information Sensitivity Perception

Existing works examined how users perceive the sensitivity of information. Markos et al. (2017) explored how users in different countries, age groups, and with varying levels of perceived privacy control perceive the sensitivity of information and their willingness to share specific information types. Their findings revealed that, while US users and Brazilians had similar rankings for specific types of information sensitivity, Brazilians were more inclined to share information compared to US users. (Schomakers et al. 2019) assessed the sensitivity of 40 types of data in a German cohort. The resulting sensitivity ranking closely resembled the one observed by Markos et al. (2017). While slight variations in the perception of information sensitivity for certain types of data were observed (Almotairi and Bataineh 2020), the overall ranking was comparable to the previous studies. Alemany et al. (2020) examined users' willingness to share information content and identified information data types that users regretted sharing. Similarly, Reinhardt et al. (2015) conducted a study with 42 Germans to evaluate the sensitivity of 20 types of content to be shared on SNS. These studies collectively indicate a consensus regarding the perceived sensitivity of different data types across various nations and cultures. However, their focus has primarily been on either the overall sensitivity of all data types (Markos, Milne, and Peltier 2017; Schomakers et al. 2019; Almotairi and Bataineh 2020) or on sensitive information shared on social networking platforms (Alemany Bordera, Del Val Noguera, and García-Fornes 2020; Reinhardt, Engelmann, and Hollick 2015). In contrast, we explore the sensitivity prediction of images using ML solutions.

Image Privacy Datasets and Taxonomies

PicAlert. Zerr et al. (2012) introduced PicAlert originally composed of 37,535 images annotated as either private or public by 81 participants aged between 10 to 59. The images were crawled from publicly shared Flickr images over a period between January and April of 2010. As in our study, they were directed to imagine that they had captured the images themselves. They then had to categorize them as either private, public, or undecidable. In the instructions, the images were defined as private, when they belonged to the private sphere, e.g., selfies, images with family members, friends, or their own interiors, or contained information that is not intended to be shared with others, such as confidential documents. All remaining images were to be classified

as public. If the participants disagreed, the images were presented to a larger group until a consensus was reached. Ultimately, only the images that were labeled as either public or private were further considered for sensitivity prediction.

PrivacyAlert. (Zhao et al. 2022) crawled 20,000 Flickr images, 83% of them were posted from 2015 to 2021 and the rest from 2011 to 2015. The images were filtered based on ten defined privacy taxonomy categories, such as nudity/sexual or unorganized home, and the corresponding keywords (see below). The authors did not indicate the exact number of participants. The participants were asked to label images according to either *clearly private, private, public, or clearly public* classes. Half of the images were annotated three times by three different annotators and used for training; the other half were annotated five times and used for validation and testing. The dataset is divided into *private* and *public*.

Image Privacy Taxonomies. Orekondy et al. (2017) outlined attributes for private image description combined with user preferences to estimate privacy exposure. However, the solution can not be utilized to model image privacy prediction in *private* or *public* classes. Li et al. (2018) defined sensitivity categories (e.g., identity, nudity) to classify images, based on a literature review. Li et al. (2020) generated a taxonomy based on existing literature and a user study that recorded users' sharing preferences. Zhao et al. (2022) adopted Orekondy et al. (2017) and Li et al. (2020) privacy taxonomies when generating PrivacyAlert. Since we are leveraging PrivacyAlert images as basis for our own dataset, we indirectly include these solutions in our work, but we go beyond the definition of such taxonomies.

ML-based Image Sensitivity Prediction

Generic Models using DL Models on PicAlert or PrivacyAlert. (Zerr et al. 2012) proposed a classification scheme based on SVM model. They used visual (SIFT and faces) and textual (tags, title) features. Their classification was based on 4,701 private and 4,701 public images. Tonge et al. (Tonge and Caragea 2016, 2018, 2020) then proposed alternative approaches, which all used both deep features derived from DL models and user tags as feature representations. However, Tonge et al. (2016) extracted deep features using a CNN model, with user tags of those images, while they investigated pre-trained AlexNet CNN for deep features (Tonge and Caragea 2018) and compared AlexNet, GoogleNet, VGG-16, and ResNet pre-trained models on object recognition in (Tonge and Caragea 2020). The models were then fine-tuned to predict image sensitivity. Besides the image content, they analysed user tags using SVM and text-based CNN in (Tonge and Caragea 2020). They used SVM for the classification into private or public classes. In contrast, Zhao et al. (2021) fine-tuned BERT to model images based on their user tags. (Zhao et al. 2022) investigated the image privacy prediction by fine-tuning ResNet pre-trained models on object and scene recognition on PrivacyAlert. The scene recognition was also examined in a former work by Tonge et al. (2018). Moreover, Zhao et al. (2022) fine-tuned BERT with user-tags only, as well as, user and deep features together. They also utilized Gated Fusion as a multi-modal approach, which included both images and user-based tags. A multi-modal approach was already considered by Tonge et al. (2019) but PicAlert was used as a benchmark instead.

In contrast, we hence investigate the performance of recent DL models, i.e., BEiT, EVA-02, ConvNeXt V2 and ALBEF, using a novel dataset based on a German cohort. We also investigate the performance of generic, user-wise, and hybrid BEiT fine-tuning on each participant.

Personalized-Based Models on PicAlert. To address the limitations of generic consensus-based models, Xioufis et al. (2016) and Zhong et al. (2017) investigated user-wise image sensitivity prediction. The former proposed a hybrid model combining user-based samples and generic image annotations, comparing it with a user-wise model. However, their method suffers from the following limitations: 1) a small user study size of only 27 participants, 2) data source inconsistencies (i.e., utilizing annotated image samples from one cohort, such as PicAlert, and assessing them with a different cohort has shown to inaccurately represent privacy trends, resulting in a significant performance decrease (Zhao et al. 2022), and 3) outdated deep learning techniques. Similarly, Zhong et al. (2017) introduced a statistical approach based on user demographics and their annotations to categorize users by privacy groups. Their user study included 114 participants. Their focus was, however, on user grouping rather than image sensitivity prediction. They argued that their personalized models have, however, disadvantages. For example, they have (1) a poor performance, due to limited user-based training data and (2) a high computational complexity of the training step required for user-based models.

Summary. We are the first to the best of our knowledge to (1) consider a solely German-speaking cohort in evaluating the image sensitivity to be shared in SNS, (2) extensively address the differences between fine-tuning and evaluating between an isolated country cohort against other cohorts, and (3) fine-tune BEiT, EVA-02, and BERT with EVA-02 features for image privacy prediction under both generic and personalized modellings. By doing so, we further address several limitations of existing works, not only on the adopted models, but especially also on the choice of a diverse age groups and gender cohort corresponding to SNS usage, a significantly higher number of participants and/or of images.

Data Collection Methodology

Recall that our goal is to assist users in better protecting their privacy by alerting them if the image they are about to post in SNS is considered as sensitive. To this end, we hence need to learn which images are sensitive for each user using labelled images. Since the available datasets suffer from different limitations (see related work), we have created a new dataset by leveraging an online questionnaire-based study according to the following methodology. Note that we plan to publicly release the dataset in the future.

Study Design

Participants were first informed about the study, the data collection, and data handling process according to the current

data protection regulations. In the study presentation, they were explicitly instructed to think as "if [they were] hypothetically, the ones that took the pictures and [were] in a position to share those images". After their explicit consent, they started to annotate 60 images allocated to them. To this end, we have designed, implemented, and deployed a web annotation tool for image annotation. Half of the images are selected from the private images of either PicAlert or PrivacyAlert; the other half are randomly selected in public images of the same datasets. The annotation tool was hosted on our institution servers. We have implemented measures to safeguard the participants' privacy, e.g., by anonymizing the collected data. Below each image, the participants were presented with the following statement: "*I find this picture sensitive with regards to my privacy*". They had to decide between *strongly disagree, disagree, agree, strongly agree*, and *I don't know*. Note that they were asked to choose with whom they would share the image in a follow-up question. Lastly, they fulfilled a questionnaire about their demographics, SNS usage, and interpersonal relationship preferences.

Image Selection Strategy. Our image crawling strategy is based on both existing datasets: PicAlert and PrivacyAlert introduced in related work. We first scraped the available images from the above datasets. We have retrieved 29,204 PicAlert images, incl. 6,676 private ones. We selected all private images and randomly selected the same number of public images. We assume that selecting a balanced number of sensitivity classes would lead to a rather balanced set of images labelled as sensitive and non-sensitive also in our study. By doing so, we solve the problem of obtaining a majority of images labelled as public. Similarly, we have scraped and retrieved PrivacyAlert images, including 1,513 private images of PrivacyAlert. 191 of them were inaccessible due to deletion or being set as private by their authors.

The motivation to utilize PicAlert and PrivacyAlert images is two-fold. Firstly, selecting other images without previous verification may have lead to a bias towards public images, since publicly accessible platforms like Flickr are more likely to contain public images as confirmed in prior datasets. Secondly, using the same images allows us to compare our results with these baselines and explore variations in privacy perception among distinct cohorts.

Ethical Aspects

To crawl and collect these images, we use the "Public Domain Dedication" and "Public Domain Mark" licenses. As a result, our dataset only includes pictures under a public license. Our dataset is based on PicAlert and PrivacyAlert selected images. We have collected all the images of PicAlert and PrivacyAlert and have employed the corresponding image user tags from PicAlert and PrivacyAlert repositories. Any collected image is prone to potentially identifiable personal metadata. While our institution does not have a formal IRB process, we have ensured to minimize potential harms from our study by respecting the Code of Ethics and the Standards of Good Scientific Practice. We have obtained approval for our user study including its questionnaire from our data protection officer. We have informed the participants about the responsible person for the data collection and handling (as defined by Art. 4 No. 7 EU-GDPR) and that the data will be stored for a period of 10 years in order to prove compliance with the guidelines of good scientific practice upon request. Our participants were sourced from the Respondi panel provider that adheres to the ISO 20252-2019 standards (ISO 2019) thus ensuring the quality of the provided services. We have received anonymized participant IDs. The participants have received financial compensation for their time. The estimated time required to complete the study was one hour. They were allowed to take breaks.

Study Distribution

Recruited by our panel provider, 920 participants completed our study. For our analysis, we have excluded 13 of them, who labeled 50 or more images (out of 60) with the same label. While these participants might have more extreme privacy conceptions than the remaining ones, they may have rushed through the questions. We hence consider the 907 remaining participants in what follows. Their age distribution is in line with the average distribution of both Facebook and Instagram users in Germany (NapoelonCat 2022a,b).

Resulting SensitivAlert Dataset

Among these participants, 670 have annotated PicAlert images and 237 PrivacyAlert images. To differentiate between the origin of the images and the resulting annotations, we have created two separate subsets. The first subset, referred to as $SensitivAlert_{Pic.}$, consists of images from the PicAlert dataset that were re-annotated by our cohort. The second subset, $SensitivAlert_{Priv.}$, consists of images from the PrivacyAlert dataset that were re-annotated by our cohort. Each image has been annotated three times in $SensitivAlert_{Pic.}$ dataset resp. up to five times in $SensitivAlert_{Priv.}$. This approach ensures a reliable assessment of our generic-based models, as we utilize $SensitivAlert_{Priv.}$ for evaluating our models on the validation and test sets. Conversely, regarding the images used for fine-tuning ($SensitivAlert_{Pic.}$), having three annotators for each image (instead of e.g., five), allows us to cover a wider variety of annotated images. For each dataset, we first merge the four sensitivity classes to the two public and private classes as also adopted in (Zhao et al. 2022). We further create a consensus dataset by including images annotated with the majority label, i.e., at least two out of three resp. up to three out of five annotators. These datasets are further referred to as $SensitivAlert_{Pic.\bigstar}$ resp. $SensitivAlert_{Priv.\bigstar}$.

Inter-rater Agreement. Based on (Zhao et al. 2022), we estimate the *Pairwise Agreement* (PA) by averaging the pairwise distance between individuals' annotations for each of the images, i.e., weighting strongly sensitive, sensitive, non-sensitive, and strongly non-sensitive with 0.00, 0.25, 0.75 and resp. 1.00. We do not include the 'I do not know' annotation in our calculation. We reach a PA of 0.65 for $SensitivAlert_{Pic.\bigstar}$ resp. 0.64 for $SensitivAlert_{Priv.\bigstar}$, indicating moderate agreement between annotators (Zhao et al. 2022). Note that the PA was 0.81 in (Zhao et al. 2022).

Age	$SensitivAlert_{Priv.}$	$SensitivAlert_{Pic.}$	Both
18-24	57	138	195
25-34	65	185	250
35-44	50	140	190
45-54	34	95	129
55-64	25	71	96
65-67	6	41	47

Table 1: Age distribution in our sample

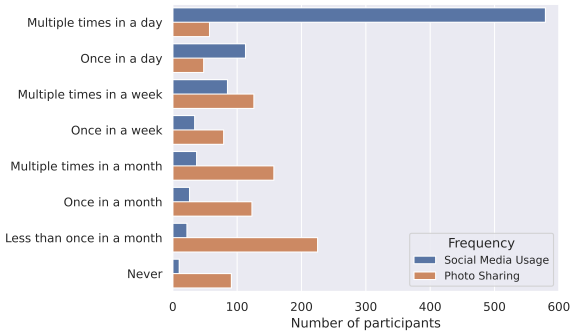


Figure 1: SNS usage and photo sharing frequency of $SensitivAlert$ participants.

	$SA_{Pic.\bigstar}$	PicAlert	$SA_{Priv.\bigstar}$	PrivacyAlert
Private	5672	7518	1156	1704
Public	6431	24615	1461	5096
Overall	12103	32133	2617	6800

Table 2: Comparison between the number of labelled images in our SensitivAlert (SA) subsets and the prior datasets.

Such a difference between annotators may be due to our wide demographic range of our participants.

Demographics and Sensitivity Distribution.

$SensitivAlert_{Pic.}$ **and** $SensitivAlert_{Pic.\bigstar}$**.** Among the 670 participants, 332 are male, 332 are female, and six are diverse. Tab. 1 shows the age distribution. Most participants use one of the SNS multiple times a day, but do not share images as often, as shown in Fig. 1. Three participants annotated each image in $SensitivAlert_{Pic.}$. Fig. 2a highlights the distribution of annotations per participant. In our consensus dataset $SensitivAlert_{Pic.\bigstar}$, we discard images that did not receive at least two out of three identical labels. This results in 5,672 sensitive and 6,431 non-sensitive. We exclude the images that reached the 'I don't know' consensus.

$SensitivAlert_{Priv.}$ **and** $SensitivAlert_{Priv.\bigstar}$**.** The age range of the 237 participants is 18 to 67, as shown in Tab. 1. 127 are male, 109 are female, and one is diverse. Fig. 2b shows the distribution of their annotations. While $SensitivAlert_{Priv.}$ contains all images annotated by each participant, $SensitivAlert_{Priv.\bigstar}$ is the consensus dataset including 1,156 sensitive images, 1,461 non-sensitive images, as displayed in Tab. 2.

Baseline Modeling

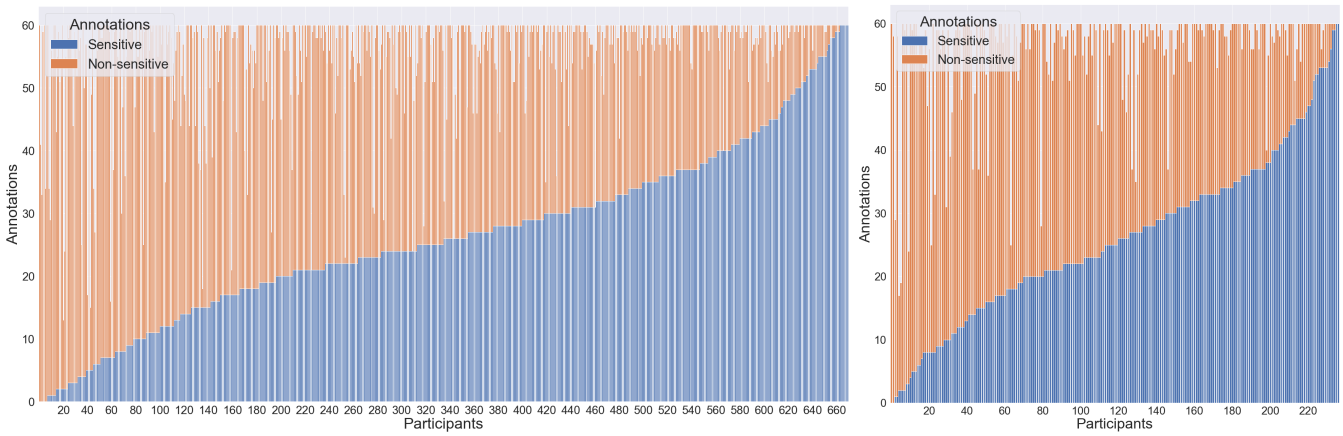
We estimate the sensitivity prediction on *SensitivAlert* using the subsequent single-modal (i.e., either visual or textual only) models: (1) objects derived from image content alone and (2) image tags based on user, deep features or both. Moreover, (3) we consider a multi-modal (i.e., visual and text jointly) model. We also investigate the role of particular privacy taxonomies on overall object-based predictions.

Object-based Privacy Prediction. Objects in images can help in distinguishing sensitive from non-sensitive images. For instance, a bathtub may be an indicator for a private image. We thus fine-tune pre-trained BEiT model introduced by Bao et al. (Bao et al. 2021) models on object classes of ImageNet-1k, along with a linear layer as a classifier head. BEiT (Bao et al. 2021) is a regular *Vision Transformers* (ViT) model (Dosovitskiy et al. 2020) pre-trained in a self-supervised setting. Commonly, these models are evaluated using ImageNet-1k (Russakovsky et al. 2015) as a benchmark, among others, along with their pre-trained weights being released. The ImageNet-1k dataset contains 1,000 object categories with 1,281,167 training images, 50,000 validation images, and 100,000 test images (Deng et al. 2009). BEiT fine-tuned and evaluated on ImageNet-1k achieved a high top-1 accuracy of 86.3% (i.e., the models ability to accurately predict an image's object class by selecting the most probable predicted class). In contrast, EVA-02 (Fang et al. 2023) was pre-trained first on 38 million images, followed by ImageNet-22k and lastly on ImageNet-1k. It reached a top-1 of 90.6% respectively a top-5 of 99.04%. We further adopt ConvNeXt V2 (Woo et al. 2023) pre-trained on ImageNet-1k which achieves up to 88.9% depending on the size variant. Due to their high performance, we utilize those models for fine-tuning in the sensitivity prediction task.

Image Tag-based Privacy Prediction. Image tags are composed of 1) user tags that individuals employ to describe the content of a shared image and 2) deep feature words generated from object categories from DL models. We generate the top-10 object categories by using EVA-02 rather than e.g. ResNet, due to the higher performance on ImageNet-1k. We then conduct separate analyses using BERT by investigating the performance of (1) EVA-02 top-10 deep features only, (2) user features only, and (3) the combination of both. We choose BERT for classification due to its demonstrated effectiveness in image privacy classification (e.g. (Zhao et al. 2022)) and how well it aligns with constrained resources, such as available in mobile devices (Sun et al. 2020).

Multi-modal Privacy Prediction. To estimate the privacy prediction performance of images and user tags jointly, we also fine-tune ALBEF multi-modal model (Li et al. 2021) along with a linear layer. ALBEF jointly encodes the text (i.e., user tags) using BERT and the images using ViT.

Taxonomy-based Privacy Prediction. We also leverage the privacy taxonomy introduced in (Zhao et al. 2022). It contains ten categories, such as *other people, violence, medical*. Each category contains different keywords. For example, the category "violence" includes "guns", "war", and "firearms weapons". If any of these keywords correspond



(a) $SensitivAlert_{Pic.}$ (b) $SensitivAlert_{Priv.}$

Figure 2: Distribution of private and public labels in the $SensitivAlert$ dataset per participant sorted by ascending number of sensitive images. Missing labels correspond to images annotated as "I don't know".

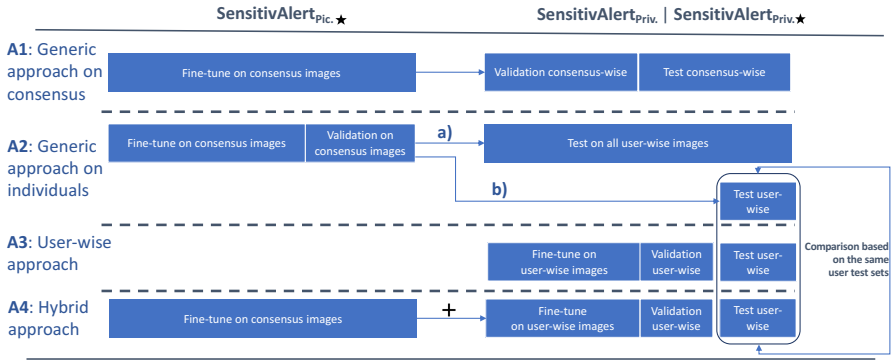


Figure 3: Pipeline of our main approaches. "Test user-wise" bar corresponds to 25% of the participants' image samples.

to a user tag associated with a $SensitivAlert_{Priv.\bigstar}$ image, we attribute that specific image to its corresponding taxonomy categories. Each category can have images annotated as public or private. Thus, we can investigate the contribution of each image category on predicting the image sensitivity.

Prediction Approaches

We utilize the introduced baseline models on several prediction approaches. We differentiate between generic fine-tuning (A1-A2) and personalized fine-tuning approaches (A3-A4), as well as, the evaluation based on the consensus (A1) and individual image annotations (A2-A4) as outlined in Fig. 3. This allows us to determine the most suitable approaches not only for the general privacy perception of SNS users, but also on each of them individually. Both generic approaches (i.e., A1 and A2) are based on consensus fine-tuning only. Specifically, A1 is based on fine-tuning of single object-based or image-tags based classification models and a multi-modal (i.e., with user tags and images inputs jointly) model using $SensitivAlert_{Pic.\bigstar}$. They are validated and evaluated on the other consensus set, i.e., $SensitivAlert_{Priv.\bigstar}$ to estimate the sensitivity prediction performance based on image content only and/or on associated tags. A1 is further used to investigate the modelling based on different cohorts/benchmarks. A2 is based on training and validation of a single object based (BEiT) classification model on consensus images of $SensitivAlert_{Pic.\bigstar}$. We evaluate A2 in two modes, (A2a) testing on all user-wise images of $SensitivAlert_{Priv.}$, and (A2b) testing on a subset of user-wise images, i.e., the same test subset as the one in user-wise (A3) and hybrid (A4) approaches. By doing so, it allows us to directly compare the performance of generic approach based on all images of an individual (i.e., A2a mode) and the performance differences between generic approach evaluated on individuals (A2b), user-wise (A3), and hybrid (A4) approaches on the same users' test sets. In user-wise (A3) and hybrid (A4) approaches, the models are fine-tuned, validated, and tested for each user. The user-wise (A3) approach is fine-tuned on user training image subset only, whereas the hybrid (A4) combines the user training image subset with $SensitivAlert_{Pic.\bigstar}$. Thus, user-wise (A3) strength relies solely on user-specific personal preferences, whereas the hybrid (A4) additionally makes use of observed general patterns from consensus labelled images. Particu-

Models	Private (%)			Public (%)			Overall (%)			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Acc.	Prec.-weighted	Rec.-weighted	F1-weighted
BEiT	70.85	82.70	76.31	84.39	73.24	78.40	77.42 (0.0095)	78.44 (0.0087)	77.42 (0.0095)	77.48 (0.0098)
EVA-02	69.77	86.09	76.69	86.59	70.32	77.16	76.88 (0.0087)	79.15 (0.0087)	77.31 (0.0078)	77.32 (0.0080)
ConvNeXt V2	69.85	79.63	74.40	81.74	72.57	76.86	75.70 (0.0046)	76.48 (0.0055)	75.70 (0.0046)	75.77 (0.0047)
BERT (UT)	64.79	80.44	71.72	81.21	65.72	72.56	72.18 (0.0070)	74.02 (0.0044)	72.18 (0.0070)	72.20 (0.0082)
BERT (eva02:DF)	65.72	79.72	72.03	80.90	67.39	73.51	72.80 (0.0141)	74.26 (0.0111)	72.80 (0.0141)	72.87 (0.0143)
BERT (UT+eva02:DF)	66.17	82.21	73.30	82.51	66.58	73.67	73.49 (0.0068)	75.29 (0.0083)	73.49 (0.0068)	73.51 (0.0069)
ALBEF	67.87	82.47	74.44	83.34	69.15	75.56	75.02 (0.0065)	76.53 (0.0024)	75.02 (0.0065)	75.06 (0.0067)

Table 3: Results obtained for single-modal (1) object-based (BEiT, EVA-02, and ConvNeXt V2), (2) image tag (BERT) approaches (i.e., *User Tags* (UT), *Deep Features* (eva02:DF), and both (UT+eva02:DF), as well as, (3) multi-modal (i.e., UT and images jointly) with ALBEF: fine-tuned and evaluated using SensitivAlert. Averages and standard deviation are over five runs.

larly, for A2b, A3, and A4, we randomly split the images labelled by the same user in three parts, i.e., 50% for training, 25% for validation, and the remaining 25% for testing.

Experimenting Details. We run all of the experiments five times each. We have used our institution's cluster to run our models. We run our code using Slurm scheduler on several nodes with hardware configurations varying from NVidia GTX 1080, GTX 980, Quadro RTX5000, and Tesla V100/32G nodes. We use BEiT-large pre-trained with images rescaled to the 224x224 pixels. We use a learning rate of $2e^{-5}$, train batch size of 10, evaluation batch size of 5, 15 epochs per training, and a weight decay of 0.01. We train ConvNeXt V2 (tiny) on the same parameters as "BEiT-large" except changing the evaluation batch size to 4 and the number of training epochs to 25. We utilize "eva02" large pre-trained with a batch size of 4, 8 epochs, and images rescaled to the 448x448 pixels. For BERT model, we use "BERT-base-uncased" pre-trained with a learning rate of $2e^{-5}$, batch size of 32, adam epsilon of $1e^{-8}$, 4 training epochs, and a maximal sequence length of 384. We train ALBEF in a batch size of 16, over 20 epochs, with a learning rate of $1e^{-5}$, and a max. sequence length of 384.

Results

We next present the results obtained by applying the approaches and models described in the previous sections.

Generic Approach Evaluated on Consensus (A1)

We first evaluate the performance of the models in our own cohort. We then analyse the cohort differences for the annotated images, before we switch between fine-tuning our models in one cohort and evaluating them in another. We finally investigate the consensus-oriented fine-tuning performance across various categories of taxonomy images.

Privacy Prediction in $SensitivAlert$ based on Objects and Image Tags. We first investigate the performance of the models based on objects and image tags by training on SensitivAlert$_{Pic.\bigstar}$ and evaluating on SensitivAlert$_{Priv.\bigstar}$. We obtain the best results with object-based fine-tuned BEiT model followed by EVA-02 and ALBEF, outperforming all of the image tag approaches, as presented in Tab. 3. Our results suggest that a generic approach based on objects is better suited for predicting the image sensitivity than image tags. As expected, a combination of user and deep features outperform both of them when considered alone. Our results hence show that, by using user-tags only, our model can accurately identify image sensitivity content to a certain degree, however, not as good as the models based on either the combination of user and deep features or fine-tuned on objects. Moreover, ALBEF multi-modal classification (i.e., jointly encoding user tags and image pairs) do not perform as good as BEiT. This can be attributed to the higher performance of BEiT on object recognition rather than ViT (which ALBEF uses for images). We also compare the results obtained on our SensitivAlert dataset to the ones from the original PicAlert resp. PrivacyAlert. We obtain better results with BEiT for the private class than (Tonge and Caragea 2020; Zhao et al. 2022), as shown in Tab. 4. Our more balanced score between f1-overall and f1 private can be attributed to our balanced dataset between private and public images. In contrast, our f1 is lower for the public class in all cases. This difference may be due to the significantly smaller number of public images that we used, in comparison to PicAlert and PrivacyAlert. However, we aimed to have as many images annotated as private and a roughly equal proportion of images annotated as public, leading us to a balanced modelling.

Source	PicAlert	PrivacyAlert	SensitivAlert★	Model	F1-Private	F1-Public	F1-All
Tonge et al. (2020)	✓			ResNet	0.717	0.920	0.872
Zhao et al. (2022)		✓		Gated Fusion	0.750	**0.921**	**0.878**
			✓	BEiT	0.763	**0.784**	**0.775**
			✓	EVA-02	**0.767**	0.772	0.773

Table 4: Best performing models of different approaches using different benchmarks.

Analysis of Cohort Differences for Annotated Images. We first assess the correlation of each annotated image between our cohort in relation to PicAlert and PrivacyAlert using an Intra-class Correlation Coefficients (ICC) model,

(a) Public in PicAlert - Private in SensitivAlert$_{Pic.\bigstar}$ (b) Private in PicAlert - Public in SensitivAlert$_{Pic.\bigstar}$ (c) Public in PrivacyAlert - Private in SensitivAlert$_{Priv.\bigstar}$ (d) Private in PrivacyAlert - Public in SensitivAlert$_{Priv.\bigstar}$

Figure 4: High frequency user tags and deep features clouds from images annotated differently between cohorts.

i.e., ICC(3,k). This assesses each image by each consensus of raters, where the consensus ratings are the only raters of interest. We used the reliability calculated by taking an average of the two consensus ratings' measurements. ICC(3,k) in cross-consensus ratings is 0.62 between our cohort in relation to PicAlert resp. 0.69 to PrivacyAlert for the same annotated images. These results show that a good correlation across cohorts exist. To identify categories of images in which the differences lie, we plot 50 most frequent user tags and deep features for each class difference in comparison to both of the existing datasets as shown in Fig. 4. In Fig. 4a, we observe "bridge", "movie", "sand", "coast", and "theater" categories among others. This shows that our cohort perceives such images as private, contrary to the cohort of PicAlert. Similarly, "swimming", "bathing", "trunks", and "malliot" are categories where our cohort considers them as public, differently from PicAlert cohort, as shown in Fig. 4b. We identify additional categories between our cohort and PrivacyAlert. Our participants perceive "tobacconist", "shop", "bulletproof", "party", "miniskirt" etc. as private (Fig. 4c) and "public domain", "brassiere", "bra", "jacket", etc. as public (Fig. 4d), contrary to PrivacyAlert cohort.

Training on $SensitivAlert_{Pic.\bigstar}$ - Evaluation on PrivacyAlert and Vice Versa. To investigate the potential differences between the performance of the models fine-tuned and evaluated across different cohorts used as benchmarks, we measure the performance of the models by Ⓐ training on $SensitivAlert_{Pic.\bigstar}$ and testing on PrivacyAlert against Ⓑ the training on original PrivacyAlert and evaluating on $SensitivAlert_{Pic.\bigstar}$. BEiT performance of Ⓐ significantly outperforms Ⓑ by 22.41% (Tab. 5). This demonstrates the high impact of the choice of the cohorts. Moreover, training with our dataset allows to better recognise the privacy patterns in other cohorts. This can be attributed to our selection of diverse participants in terms of SNS usage, age, and gender among others, as well as, to a larger number of images balanced between private and public images.

Training on $SensitivAlert_{Priv.\bigstar}$ - Evaluation on PicAlert and Vice Versa. We further investigate if the same findings hold when comparing between Ⓒ training on $SensitivAlert_{Priv.\bigstar}$ and testing on PicAlert against Ⓓ training the model on PicAlert annotations and evaluating on $SensitivAlert_{Priv.\bigstar}$. Ⓒ outperforms Ⓓ (Tab. 5), suggesting that our dataset is more suitable for fine-tuning models on image sensitivity due to a more diverse cohort. The

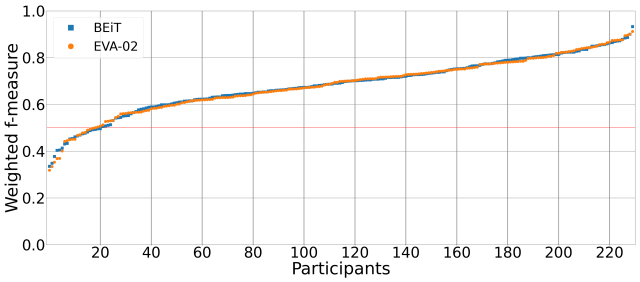


Figure 5: The generic approach's performance by fine-tuning BEiT resp. EVA-02 on $SensitivAlert_{Pic.\bigstar}$ and evaluating it on all the participant's individual image samples of $SensitivAlert_{Priv.}$ averaged over five random runs.

Train	Evaluation	Model	F1 (%)		
			Priv.	Pub.	All
$SA_{Pic.\bigstar}$	PrivacyAlert	BEiT	59.97	84.81	**79.03**
		EVA-02	57.81	77.44	72.80
		ConvN. V2	58.88	81.89	76.58
PrivacyAlert	$SA_{Pic.\bigstar}$	BEiT	37.99	72.73	56.62
		EVA-02	44.27	73.40	59.77
		ConvN. V2	46.05	73.01	**60.42**
$SA_{Priv.\bigstar}$	PicAlert	BEiT	56.92	89.91	82.28
		EVA-02	54.80	88.53	80.74
		ConvN. V2	61.19	89.49	**82.98**
PicAlert	$SA_{Priv.\bigstar}$	BEiT	70.13	77.21	**74.00**
		EVA-02	68.74	77.32	73.57
		ConvN. V2	68.48	76.80	73.13

Table 5: Our object-based model results obtained by fine-tuning and evaluating on different cohorts.

difference is lower than when comparing Ⓐ with Ⓑ. This can be attributed to the larger number of images in PicAlert as compared to PrivacyAlert.

Moreover, our BEiT model performs worse when trained and evaluated on our own dataset, i.e., reaching 77.48% in contrast to when being evaluated on other benchmarks with resp. 79.03% on PrivacyAlert and 82.28% on PicAlert (Tab. 5). The same findings hold for ConvNeXt V2 results. This further demonstrate that our models in general, although reaching lower values than others when trained and evaluated on own benchmarks (Tab. 4), ours reflect better the users image privacy in contrast to the other privacy datasets.

Taxonomy-based Privacy Prediction. We fine-tune and validate BEiT in $SensitivAlert_{Pic.\bigstar}$ and evaluate in each

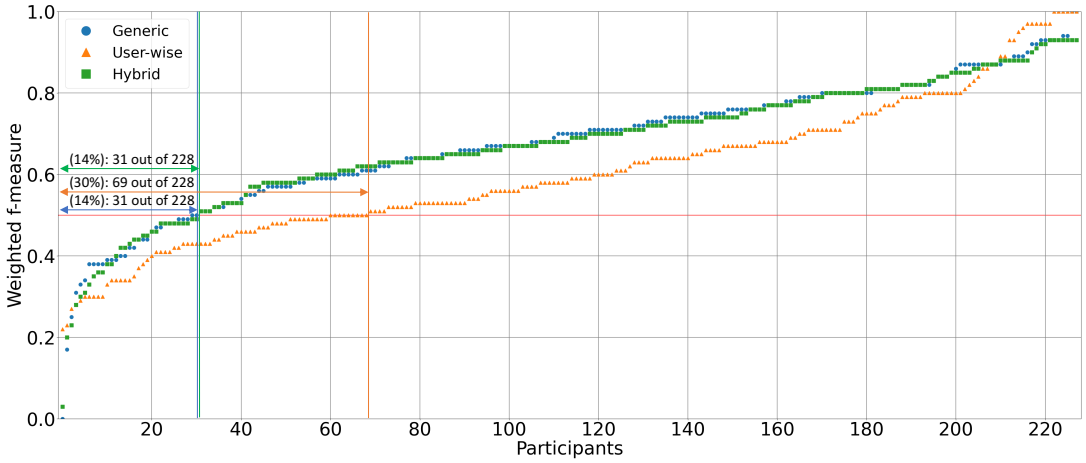


Figure 6: Performance of generic, user-wise and hybrid BEiT approaches on $SensitivAlert_{Priv.}$ participants in an ascending order of the corresponding approach. The weighted f1 results are averages over five random runs.

Evaluation category	F1 (%)		
	Private	Public	All
$SensitivAlert_{Priv.\bigstar}$	75.04	76.37	75.78
Nudity/sexual	**84.39**	59.43	76.65
Other people	76.88	58.53	68.51
Unorganized home	79.26	77.19	**78.20**
Violence	69.75	80.36	76.25
Medical/blood	69.39	78.96	75.51
Drinking/party	71.88	60.14	65.85
Appearance/facial expression	79.47	43.17	68.35
Bad Character/unlawful criminal	58.93	68.16	64.26
Religion/culture	54.38	**83.10**	76.55
Personal information	42.36	82.68	73.53

Table 6: BEiT privacy prediction performance of each privacy taxonomy category in F1. F1-overall is in weighted average. An image can be part of more categories.

	Weighted f1			
	Generic on all	Generic	User-wise	Hybrid
Eval. on	$SA_{Priv.}$		$SA_{Priv.}$	
Min	0.31	0.00	0.18	0.00
Q1	0.61	0.59	0.49	0.58
Mean	0.69	0.68	0.61	0.67
Q2	0.70	0.70	0.59	0.67
Q3	0.79	0.80	0.72	0.80
Max	0.98	1.00	1.00	0.94

Table 7: Extrema, quartiles, and mean for the generic BEiT-based approach evaluated on all participants' images and for the generic, user-wise and hybrid ones on the same test set (i.e., 25%) of each participants' images.

privacy taxonomy categories of $SensitivAlert_{Priv.\bigstar}$ to investigate which categories are contributing the most and the least in the pattern recognition of image sensitivity. As shown in Tab. 6, we observe that images characterized by *unorganized home* user tags are overall the easiest to be predicted followed by the *nudity/sexual* category. In contrast, the lowest performance for the overall estimation is obtained for the *bad character and unlawful criminal* category which has the smallest sample size. Our results imply that either a balanced sample of categories or a larger size contribute more to the image sensitivity prediction.

Generic Approach Evaluated on Participants (A2)

We next investigate the performance of generic object-based approach for individuals. Recall that we first fine-tune BEiT and EVA-02 pre-trained models using the data from $SensitivAlert_{Pic.\bigstar}$ The fine-tuned models are then evaluated on the images from the other dataset labelled by each participant, i.e., $SensitivAlert_{Priv.}$. The results in Fig. 5 show that A2 performs well for most participants. BEiT only slightly outperforms EVA-02. This demonstrates that both our BEiT and EVA-02 generic modellings accurately predict the individual privacy preferences for most SNS users.

Comparison of Generic (A2), User-wise (A3), and Hybrid (A4) Object-based Approaches

We further investigate the performance of generic, user-wise and hybrid approaches evaluated under the same participant image samples. Our goal is thus to compare and determine which is better suited for detecting sensitive images at participant level. As illustrated in Fig. 6, we observe that the hybrid and generic approaches outperform the user-wise approach for most participants. This is confirmed in Tab. 7. We further investigate the performance difference between A2b, A3, and A4 for the same participants and labelled images illustrated in Fig 7. For only 32% resp. 30% of our participants, A3 outperforms A4 resp. A2b. This implies that the A4 and A2b are more suitable for the prediction of the content sensitivity. We attribute the higher performance of them to the higher number of labelled data and the consensus fine-tuning with $SA_{PicAlert\bigstar}$. Against our expectations, A4 overall do not result in significantly better results than A2b, as shown in Fig. 7 and Tab. 7. This implies that more images per participants would be needed to investigate the

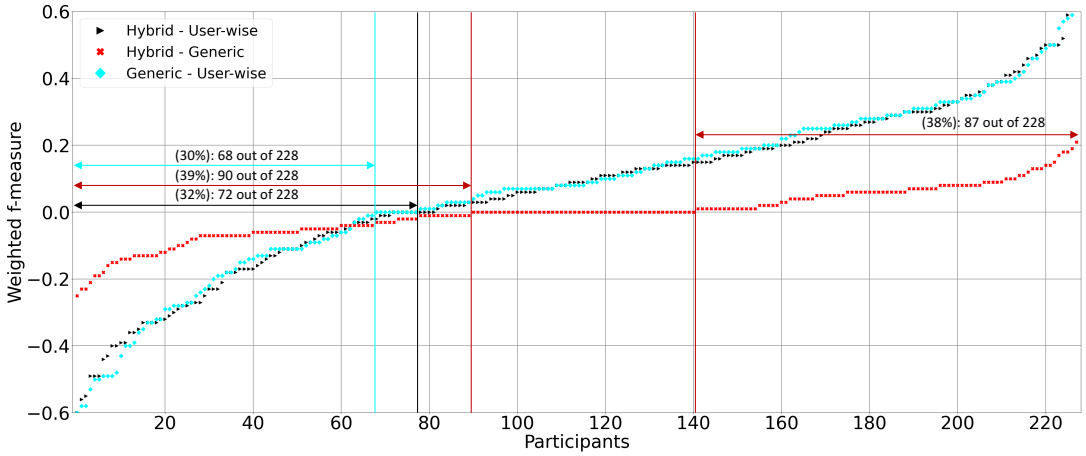


Figure 7: Performance differences of generic, user-wise and hybrid BEiT approaches on $SensitivAlert_{Priv.}$ users under the same participants' test image samples. The weighted f1 results are averages over five random runs.

Evaluation on	Generic	User-wise	Hybrid
18-24	65.73	59.90	65.21
25-34	66.37	57.96	65.16
35-44	67.33	62.26	67.56
45-54	70.78	59.95	71.05
55-64	73.92	63.73	74.47
65-67	63.62	75.35	61.87

Table 8: BEiT privacy prediction performance of approaches on different age groups. F1 is the average across subgroups.

Evaluation on	Generic	User-wise	Hybrid
Multiple times in a day	67.81	60.32	67.28
Once in a day	70.32	63.70	69.88
Multiple times in a week	71.15	52.89	70.24
Once in a week	51.85	58.89	69.88
Multiple times in a month	69.21	68.89	69.60
Once in a month	57.63	65.54	61.46
Less than once in a month	77.06	65.28	71.95

Table 9: BEiT privacy prediction performance assessed across SNS frequency usage. F1 is the subgroup average.

possible improvement of the hybrid approach (A4).

Comparison on age groups and SNS usage frequency. We further outline the performance of A2-A4 on users across different age groups and SNS usage frequencies. The generic (A2) and hybrid (A4) approaches perform better on older SNS users compared to younger ones with an exception on the age group between 65 to 67 years old, as shown in Tab. 8. The higher performance of user-wise (A3) approach on the particular age group suggest that the individual preferences are more prevalent than the consensus. Our results also show that A4 significantly outperforms A2 in users that use SN once per week, as shown in Tab. 9, suggesting that users' privacy patterns extracted from users own annotations are more prevalent on that particular activity group.

Summary. Using our dataset, we have investigated the performance of different generic, user-wise, and hybrid approaches. Our results show that image sensitivity can be modelled; it reaches an 77.48% f1 based on fine-tuned BEiT transformer model. Moreover, employing solely user tags enables our model to reasonably detect image content sensitivity. Nevertheless, its accuracy is lower than models combining them with deep features or object-based fine-tuned approaches. Training our best-performing model using our dataset yields to higher results when tested in existing datasets, than them being trained on their cohort and evaluated in ours. This suggests that our model overall captures better differences between cohorts and that the other models based on other cohorts are incapable of accurately predicting the image privacy of our cohort. In addition to comparing our work with existing datasets, we have investigated three image privacy prediction approaches on individual SNS users. Overall, the hybrid and generic outperform the user-wise approach for the majority of the participants.

Discussion

We next discuss the results in the context of possible integration of our models into an SNS interface along with our limitations. The performance of our generic-based image sensitivity classification model based on fine-tuning BEiT on our dataset leads to good results at both a consensus and an individual level. Upon further evaluation on the other existing cohorts, we observe that our model performs even better on them. This is due to our diverse participant selection, considering SNS usage, age, and gender, along with a balance of private and public images. Thus, we argue that our model is able to capture the general SNS privacy perceptions, and in turn, could be integrated in an SNS interface, wherein a user would be able to occasionally interact with the interface by accepting/discarding the suggested models' prediction when an image is being shared, as envisioned by Reinhardt et al. (2015). The model could be then further fine-tuned to suit users' preferences. Our results indicate that, in general, at least more than 30 images would be needed to be annotated

by the participant, to possibly profit from a hybrid approach. The latter could be integrated into a *Federated Learning* (FL) paradigm, wherein a BEiT classification global model fine-tuned first on a consensus dataset could be used at the initialization point (similar to Chen et al. (2023)) and participants' annotations could be incrementally utilized by their local models using for example FedProx FL algorithm. FedProx (Sahu et al. 2018) allows for devices with different resource capabilities and thus can be adopted. Alternatively, the user-wise approach is also promising, especially considering the sample of individual images upon which we evaluated. However, such an approach would demand available participant-wise annotated images, before being fine-tuned for image sensitivity prediction, thus reducing its usability.

Limitations. Our results are limited to the number of the images annotated by each participant. Especially, the performance of user-wise and hybrid approaches may be further enhanced with more annotated images per participant. However, to obtain them, the study should be carefully designed, as user's fatigue increases during the labelling process. While our goal is to explore another cultural cohort, we have not conducted an extensive comparison between several isolated cultures. Thus, a cultural bias may exist. Our annotation study relied on images crawled from Flick images. The images were either crawled by their posting dates (i.e., the images selected from PicAlert) or filtered based on privacy taxonomy categories (i.e., PrivacyAlert images). The images were shared publicly in Flickr from users. Although it might have been annotated from others as private, a bias of such selective and once publicly shared images may exist.

Conclusions

Reducing privacy exposure of SNS users is an important goal to achieve, as sharing user-generated images can lead to severe consequences. To contribute in reaching this goal, we have conducted an annotation study involving 907 German-speaking participants. Additionally to analyzing these annotations in isolation, we have compared them with the existing collected datasets to identify differences between modelling on our German cohort and the two other cohorts from different countries. Our fine-tuned models on our dataset in comparison to existing ones achieve better results than the existing being evaluated on ours, indicating that our models capture a broader range of user image privacy patterns. Using recent transformer-based DL models, our results also show that BEiT and EVA-02 fine-tuned models slightly outperform BERT and ALBEF that combine user tags with deep features resp. images in our dataset. Moreover, we have investigated image sensitivity prediction on individuals sharing preferences by fine-tuning BEiT and EVA-02 pre-trained model on object recognition. We also have leveraged the obtained dataset to investigate the performance of three designed approaches: generic, user-wise and hybrid. Our results show that the hybrid and generic approaches outperform the user-wise one. This especially confirms that a generic model also can lead to a high accuracy for most users. No significant difference was shown between generic and hybrid approaches.

Acknowledgments

This work is partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) referenced with the number #317687129. We thank Valerius Mattfeld for his work on the image annotation tool.

References

Alan, A.; Al-Arnaout, Z.; Topcu, A.; Zaki, C.; Shdefat, A.; and Elbasi, E. 2022. How Do Default Privacy Settings on Social Media Apps Match People's Actual Preferences? In *International Conference on Electrical and Computing Technologies and Applications (ICECTA)*.

Alemany Bordera, J.; Del Val Noguera, E.; and García-Fornes, A. 2020. Empowering Users Regarding the Sensitivity of their Data in Social Networks through Nudge Mechanisms. In *Proc. of the Hawaii International Conference on System Sciences*.

Almotairi, K. H.; and Bataineh, B. O. 2020. Perception of Information Sensitivity for Internet Users in Saudi Arabia. *Acta Informatica Pragensia*.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2021. BEiT: BERT Pre-Training of Image Transformers. *arXiv*.

Chen, H.-Y.; Tu, C.-H.; Li, Z.; Shen, H. W.; and Chao, W.-L. 2023. On the Importance and Applicability of Pre-Training for Federated Learning. In *The International Conference on Learning Representations*.

Chen, Y.; Zha, M.; Zhang, N.; Xu, D.; Zhao, Q.; Feng, X.; Yuan, K.; Suya, F.; Tian, Y.; Chen, K.; Wang, X.; and Zou, W. 2019. Demystifying Hidden Privacy Settings in Mobile Apps. In *IEEE Symposium on Security and Privacy (S&P)*.

Coopamootoo, K.; and Gross, T. 2017. Why Privacy Is All But Forgotten. *Proc. on Privacy Enhancing Technologies*.

Coopamootoo, K. P. 2020. Usage Patterns of Privacy-Enhancing Technologies. In *Proc. ACM on Computer and Communications Security*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Difallah, D.; Filatova, E.; and Ipeirotis, P. 2018. Demographics and Dynamics of Mechanical Turk Workers. In *Proc. of the ACM International Conference on Web Search and Data Mining*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv*.

Dressing, H.; Bailer, J.; Anders, A.; Wagner, H.; and Gallas, C. 2014. Cyberstalking in a Large Sample of Social Network Users: Prevalence, Characteristics, and Impact Upon Victims. *Cyberpsychology, Behavior and Social Networking*.

EU Commission. 2019. Special Eurobarometer 487a – The General Data Protection Regulation. Technical report.

Fang, Y.; Sun, Q.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. EVA-02: A Visual Representation for Neon Genesis. *arXiv preprint arXiv:2303.11331*.

ISO. 2019. ISO 20252:2019(EN) Market, Opinion and Social Research, Including Insights and Data Analytics.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Advances in Neural Information Processing Systems*.

Li, Y.; Troutman, W.; Knijnenburg, B. P.; and Caine, K. 2018. Human Perceptions of Sensitive Content in Photos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Li, Y.; Vishwamitra, N.; Hu, H.; and Caine, K. 2020. Towards A Taxonomy of Content Sensitivity and Sharing Preferences for Photos. In *Proc. of the CHI Conference on Human Factors in Computing Systems*.

Lipford, H. R.; Besmer, A.; and Watson, J. 2008. Understanding Privacy Settings in Facebook with an Audience View. In *Proc. of the Conference on Usability, Psychology, and Security*.

Lunheim, R.; and Sindre, G. 1993. Privacy and Computing: a Cultural Perspective. In *Proc. IFIP Conf. on Sec. and Control of Info. Tech. in Society*.

Markos, E.; Milne, G. R.; and Peltier, J. W. 2017. Information Sensitivity and Willingness to Provide Continua: A Comparative Privacy Study of the United States and Brazil. *Journal of Public Policy & Marketing*.

Murmann, P.; Beckerle, M.; Fischer-Hübner, S.; and Reinhardt, D. 2021. Reconciling the What, When and How of Privacy Notifications in Fitness Tracking Scenarios. *PMC*.

NapoelonCat. 2022a. Distribution of Facebook Users in Germany as of March 2022, by Age Group and Gender. https://www.statista.com/statistics/1029645/facebook-users-germany-age-gender/ (accessed in 01.2024).

NapoelonCat. 2022b. Distribution of Instagram Users in Germany as of March 2022, by Age Group and Gender. https://www.statista.com/statistics/1021961/instagram-users-germany-age-gender/ (accessed in 01.2024).

Orekondy, T.; Schiele, B.; and Fritz, M. 2017. Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images. In *IEEE International Conference on Computer Vision*.

Reinhardt, D.; Engelmann, F.; and Hollick, M. 2015. Can I Help You Setting Your Privacy? A Survey-based Exploration of Users' Attitudes Towards Privacy Suggestions. In *Proc. of the International Conference on Advances in Mobile Computing and Multimedia*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *Intern. Journal of Computer Vision*.

Sahu, A. K.; Li, T.; Sanjabi, M.; Zaheer, M.; Talwalkar, A.; and Smith, V. 2018. Federated Optimization in Heterogeneous Networks. *arXiv: Learning*.

Schomakers, E.-M.; Lidynia, C.; Müllmann, D.; and Ziefle, M. 2019. Internet Users' Perceptions of Information Sensitivity – Insights From Germany. *International Journal of Information Management*.

Snyder, P.; Doerfler, P.; Kanich, C.; and McCoy, D. 2017. Fifteen Minutes of Unwanted Fame: Detecting and Characterizing Doxing. In *Proc. of the Internet Measurement Conf.*

SocialPilot. 2023. 500+ Social Media Statistics You Must Know in 2023. https://www.socialpilot.co/blog/social-media-statistics (accessed in 01.2024).

Sun, Z.; Yu, H.; Song, X.; Liu, R.; Yang, Y.; and Zhou, D. 2020. MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices. arXiv:2004.02984.

Tonge, A.; and Caragea, C. 2016. Image Privacy Prediction Using Deep Features. *Proc. of the AAAI Conference on Artificial Intelligence*.

Tonge, A.; and Caragea, C. 2018. On the Use of "Deep" Features for Online Image Sharing. In *Proc. of The Web Conference Companion*.

Tonge, A.; and Caragea, C. 2019. Dynamic Deep Multimodal Fusion for Image Privacy Prediction. In *Proc. of The World Wide Web Conference*.

Tonge, A.; and Caragea, C. 2020. Image Privacy Prediction Using Deep Neural Networks. *ACM Trans. Web*.

Tonge, A.; Caragea, C.; and Squicciarini, A. 2018. Uncovering Scene Context for Predicting Privacy of Online Shared Images. *Proc. of the AAAI Conf. on Artificial Intelligence*.

Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I. S.; and Xie, S. 2023. ConvNeXt V2: Co-Designing and Scaling ConvNets With Masked Autoencoders. In *Proc. of the IEEE/CVF Conference on CVPR*.

Xioufis, E. S.; Papadopoulos, S.; Popescu, A.; and Kompatsiaris, Y. 2016. Personalized Privacy-aware Image Classification. In *Proc. of the ACM on International Conference on Multimedia Retrieval*.

Yates, J. 2017. From Temptation to Sextortion Inside the Fake Facebook Profile Industry. http://ici.radio-canada.ca/special/sextorsion/en (accessed in 01.2024).

Zerr, S.; Siersdorfer, S.; and Hare, J. 2012. PicAlert!: A System for Privacy-Aware Image Classification and Retrieval. *ACM Conf. on Information and Knowledge Management*.

Zerr, S.; Siersdorfer, S.; Hare, J.; and Demidova, E. 2012. Privacy-Aware Image Classification and Search. In *Proc. of the International ACM Conference on Research and Development in Information Retrieval*.

Zhao, C.; and Caragea, C. 2021. Knowledge Distillation with BERT for Image Tag-Based Privacy Prediction. In *Proc. of the International Conference on Recent Advances in Natural Language Processing*.

Zhao, C.; Mangat, J.; Koujalgi, S.; Squicciarini, A.; and Caragea, C. 2022. PrivacyAlert: A Dataset for Image Privacy Prediction. *Proc. of the International AAAI Conference on Web and Social Media*.

Zhong, H.; Squicciarini, A.; Miller, D.; and Caragea, C. 2017. A Group-Based Personalized Model for Image Privacy Classification and Labeling. In *Proc. of the International Joint Conference on Artificial Intelligence*.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. We have conducted an anonymized user study conforming to privacy norms (see Sec. "Ethical Aspects" for the other details). We collect publicly available Flickr images only and we do not distribute them.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes. We particularly have listed the summary of our contribution in the "Introduction" section.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. See Sec. "Prediction Approaches" and Sec. "Results" in particular.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see "Study Distribution" subsection in particular. Our user study cohort is balanced between males and females. Moreover, participants' age distribution is in line with the average distribution of both Facebook and Instagram users in Germany.
   (e) Did you describe the limitations of your work? Yes. The limitations are emphasized within Sec. "Discussion" under the "Limitations" paragraph.
   (f) Did you discuss any potential negative societal impacts of your work? Yes.
   (g) Did you discuss any potential misuse of your work? Yes.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. See Sec. "Ethical Aspects". Moreover, we plan to release the dataset with Flickr ids of annotated images and user study annotations of those images only.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   (b) Have you provided justifications for all theoretical results? N/A
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   (e) Did you address potential biases or limitations in your theoretical framework? N/A
   (f) Have you related your theoretical results to the existing literature in social science? N/A
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? We have included the instructions needed to reproduce the experimental results. However, we did not include the code and the data. We plan to release the dataset in the upcoming submission deadline of ICWSM Dataset section.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, in the 'Prediction Approaches' section, we have included the training details accompanied with an "Experimental Details" paragraph, wherein we have described the exact models that we used for fine-tuning along with their parameters.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, we did report the standard deviation of our results over five runs for the results with the generic (single modelling approaches), in the "Results" section (Tab. 3).
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, in the end of the 'Prediction Approaches' section, we have included the type of resources that we have used for computation.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes. See the "Prediction Approaches", "Results", and "Discussion" sections.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes. See the "Discussion" section.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes, we cite the creators all over the paper when it is applicable, but especially under the "Data Collection Methodology" section.
   (b) Did you mention the license of the assets? Yes, see "Ethical Aspects" subsection.
   (c) Did you include any new assets in the supplemental material or as a URL? No. In a future work, we plan to release the dataset with the annotations from our

user study and Flickr file name ids of the images in the ICWSM Dataset section.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see "Ethical Aspects" subsection.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see "Ethical Aspects" subsection.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? N/A.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? N/A.

6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? We have included the details about the the study information, data collection and data handling in the "Study Design" and "Ethical Aspects" subsections of "Data Collection Methodology" section. The full original text of instructions is excluded because of the length of the text.

(b) Did you describe any potential participant risks, with mentions of *Institutional Review Board* (IRB) approvals? The information is described in the "Ethical Aspects" subsection. While our institution does not have a formal IRB process, we ensured to minimize potential harms from our study by respecting the Code of Ethics and the Standards of Good Scientific Practice.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? We have included the total amount spent on participant compensation.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, see "Ethical Aspects" subsection.