Personalized Privacy-aware Image Classification

Conference Paper · June 2016

DOI: 10.1145/2911996.2912018

4 authors:

Eleftherios Spyromitros-Xioufis
Expedia
31 PUBLICATIONS **2,002** CITATIONS
SEE PROFILE

Symeon Papadopoulos
The Centre for Research and Technology, Hellas
256 PUBLICATIONS **4,720** CITATIONS
SEE PROFILE

Adrian Popescu
Atomic Energy and Alternative Energies Commission
130 PUBLICATIONS **2,084** CITATIONS
SEE PROFILE

Ioannis (Yiannis) Kompatsiaris
The Centre for Research and Technology, Hellas
1,023 PUBLICATIONS **14,035** CITATIONS
SEE PROFILE

Personalized Privacy-aware Image Classification

Eleftherios Spyromitros-Xioufis[1], Symeon Papadopoulos[1], Adrian Popescu[2], Yiannis Kompatsiaris[1]
[1]CERTH-ITI, 57001 Thessaloniki, Greece, {espyromi,papadop,ikom}@iti.gr
[2]CEA, LIST, 91190 Gif-sur-Yvette, France, adrian.popescu@cea.fr

ABSTRACT

Information sharing in online social networks is a daily practice for billions of users. The sharing process facilitates the maintenance of users' social ties but also entails privacy disclosure in relation to other users and third parties. Depending on the intentions of the latter, this disclosure can become a risk. It is thus important to propose tools that empower the users in their relations to social networks and third parties connected to them. As part of USEMP, a coordinated research effort aimed at user empowerment, we introduce a system that performs privacy-aware classification of images. We show that generic privacy models perform badly with real-life datasets in which images are contributed by individuals because they ignore the subjective nature of privacy. Motivated by this, we develop personalized privacy classification models that, utilizing small amounts of user feedback, provide significantly better performance than generic models. The proposed semi-personalized models lead to performance improvements for the best generic model ranging from 4%, when 5 user-specific examples are provided, to 18% with 35 examples. Furthermore, by using a semantic representation space for these models we manage to provide intuitive explanations of their decisions and to gain novel insights with respect to individuals' privacy concerns stemming from image sharing. We hope that the results reported here will motivate other researchers and practitioners to propose new methods of exploiting user feedback and of explaining privacy classifications to users.

1. INTRODUCTION

Uploading and sharing information in Online Social Networks (OSNs) is nowadays a frequent activity for the majority of Internet users. Such shared pieces of information are aggregated into digital user profiles. These profiles support a business model based on free access to the OSN service and users have little or no control on how their profiles are exploited by the OSN. Typically, OSNs employ sophisticated algorithms to make sense of the data posted by their users in order to create personal profiles that they often use to perform ad targeting. Different research and industrial initiatives point out risks related to different aspects of information sharing. The `Sunlight` project[1] enhances transparency by detecting which textual data are used for personalized advertising. `PleaseRobMe`[2] illustrates a straightforward implication of explicit location disclosure. Note, however, that location disclosure can also be implicit, e.g. through one's posted images. Our own contribution is part of `USEMP`[3], a multidisciplinary European project which develops DataBait[4], a tool that provides feedback about what can be inferred from a user's historical data shared on OSNs.

These above initiatives provide important contributions to understanding the risks related to information sharing. However, most of the challenges related to the proposal of effective privacy preservation tools lie ahead. First, researchers need to operate under the privacy paradox [15] which causes a discrepancy between what users intend to share and what they actually disclose. This discrepancy impedes the wide adoption of privacy preservation methods and tools and an important effort is needed to educate users toward its reduction. Second, privacy breaches might be caused by different types of disclosed data, including multimedia documents and behavioral or social links in OSNs. Research should focus on these data individually, as well as on their interlinking. Third, privacy perception is inherently subjective and dynamic. Consequently, its modeling should include a strong personalization component that caters to each user's needs. Equally important, the models should evolve over time. Fourth, effective privacy preservation tools might be perceived as a threat for current business practices that are built around the exploitation of user profiles. The adoption of such tools by OSNs is conditioned by public demand combined with regulatory requirements. The Privacy Impact Assessment required by the US E-government Act[5] issued in 2002 and the proposed European General Data Protection Regulation[6] are examples of how regulation acts upon business practices. Fifth, the proposed privacy inference methods should work under real-time constraints in order to give immediate feedback to the user, preferably before the information is shared on OSNs. These

ICMR '16, June 6–9, 2016, New York, NY, USA.

ISBN 978-1-4503-4359-6/16/06…$15.00

DOI: http://dx.doi.org/10.1145/2911996.2912018

[1]http://columbia.github.io/sunlight
[2]http://www.pleaserobme.com
[3]http://www.usemp-project.eu
[4]https://databait.hwcomms.com
[5]https://www.gpo.gov/fdsys/pkg/PLAW-107publ347
[6]http://ec.europa.eu/justice/data-protection/reform/index_en.htm

Figure 1: A hardly comprehensible justification (green rectangles highlighting the most discriminative local patches) provided for a private classification by the PicAlert system. Image from [23]

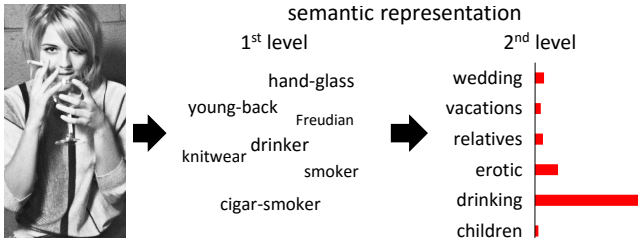

Figure 2: A better justification of the classifier's decision consisting of a tag-cloud of the most prevalent image tags and a projection of those tags into six privacy-related dimensions

methods should also offer high-quality and understandable results in order to be adopted by users. Finally, the creation and sharing of privacy-related evaluation datasets is difficult due to the very nature of the information included. However, such datasets are essential to evaluate the merits of proposed methods in a quantifiable way and facilitate reproducibility.

Here, we tackle the privacy-related risks in the context of image sharing and try to tackle some of the challenges mentioned above. Image sharing is such a widely used and valued service that preventing users from sharing their images cannot be considered as a viable means of protecting their online privacy. Instead, having access to a service that could automatically process one's images before they are shared with the OSN, and being alerted in case their content is found to be sensitive, would be a very practical and transparent way of safeguarding the online privacy of OSN users without affecting their image sharing experience.

A first solution to this problem was presented in [24], where the authors defined the problem of automatically classifying users' images as being of private or public nature, and tested the effectiveness of standard image and text features in a supervised learning setting for solving the problem. In that work, the authors focused on developing models that capture a *generic* ("community") notion of privacy, making the underlying assumption that each user perceives privacy in the same way. However, OSN users often have wildly different perceptions and norms regarding privacy [16]. A further limitation of that solution is that the classification decision was justified by highlighting the most discriminative local patches in the image as shown in Figure 1. Such a justification is hardly comprehensible by non-experts in computer vision. Providing more intuitive, higher-level, explanations would be clearly more desirable. An example is given in Figure 2, where a private classification is accompanied by an automatically generated cloud of the most prevalent image tags and a projection of those tags into a number of privacy-related dimensions.

In this paper, we propose a personalized image privacy scoring and classification system that provides an effective privacy safeguarding mechanism on top of image sharing OSN facilities and alleviates the limitations of previous solutions. In particular, we make the following contributions:

- **Personalized privacy classification:** We demonstrate that by combining feedback from multiple users with a limited amount of user-specific feedback, we can obtain significantly more accurate privacy classifications compared to those obtained from a generic model (Section 4.3).

- **Real-world dataset:** We create a realistic benchmark dataset via a user study where participants annotate their own photos as private or public according to their own notion of privacy. Experiments on this dataset reveal the limitations a generic privacy definition and highlight the necessity of building personalized privacy classification models (Section 4.2).

- **Semantic justification:** We employ a type of semantic features that facilitate the explanation of image privacy classifications and support the discovery of valuable insights with respect to users' privacy concerns (Section 4.4). Importantly, these features are computed based solely on the visual content of the images and, therefore, the approach does not require the existence of manually assigned image tags.

- **State-of-the-art performance:** By using visual features extracted from deep convolutional neural networks (CNNs) we significantly improve the state-of-the-art performance on an existing private image classification benchmark (Section 4.2).

2. RELATED WORK

Most modern OSNs allow users to control the privacy settings of their shared content. Yet, the typical user finds it difficult to understand and correctly configure the offered access control policies [12]. As a result, several studies [11, 12] have identified a serious mismatch between the desired and the actual privacy settings of online shared content. This discrepancy motivated the development of mechanisms that aid users in selecting appropriate privacy settings. In the work of [14], for instance, the authors focused on Facebook posts and evaluated prediction models that make use of users' previous posts and profile preferences in order to suggest appropriate privacy settings for new posts. Despite achieving high performance, the authors noticed differences in user behaviors and concluded that personalized privacy models could further improve the results.

Zerr et al. [24], were among the first to consider the problem of privacy-aware image classification. In their work, a large-scale user study was conducted asking participants to annotate a large number of publicly available Flickr photos as being either "private" or "public". The study was set up as a social annotation game where players were instructed

to adopt a common definition of privacy[7] and were rewarded for providing annotations that were similar to those of other players. The resulting dataset, referred to as *PicAlert*, was used to train supervised classification models that capture a generic ("community") notion of privacy.

Extending that work, [21] experimented with combinations of visual and metadata-derived features and achieved better prediction accuracy on PicAlert. [21] also attempted to solve a more complex privacy classification problem where three types of disclosure were defined for each image (view, comment, download) and the task was to assign one of five privacy levels ('Only You', 'Family', 'Friends', 'SocialNetwork', 'Everyone') to each type of disclosure. As in [24], their models captured only a generic perception of privacy.

Differently from the majority of previous works, our paper highlights the limitations of generic image privacy classification models and proposes an effective personalization method. To the best of our knowledge, [4] is the only work that considers privacy classification of personal photos as we do here. However, [4] evaluates only purely personalized models, assuming that each user provides sufficient amount of feedback. In contrast, our method achieves high performance even at the presence of very limited user-specific feedback by leveraging feedback from other users. Moreover, while [4] uses only metadata-based (location, time, etc.) and simple visual features (colors, edges, etc.), we employ state-of-the-art CNN-based semantic visual features that facilitate comprehensible explanations of the classification outputs. Very recently, [22] evaluated the performance of deep features on PicAlert (again in the context of a generic privacy model) and found that they yield remarkable improvements in performance compared to SIFT, GIST and user-assigned tag features. Moreover, the authors evaluated the performance of 'deep tag' features (which are similar to the first level of semantic features that we extract here) but did not exploit them for justifying the classifier's decisions.

3. APPROACH

3.1 Personalized Privacy Models

Privacy classifications based on a generic privacy classification model as the one developed in [24] are undoubtedly useful for preventing users from uploading images that are considered to be private according to a generic notion of privacy. However, as the perception of privacy varies greatly among users depending on factors such as age, social status and culture, it is expected that a generic model would provide inaccurate predictions for certain users, thus decreasing the reliability and usefulness of the alerting mechanism. To overcome this issue, we propose the exploitation of user feedback in order to build personalized privacy models. Such feedback could be acquired either explicitly, by asking OSN users to provide examples of private and public photos, or implicitly, by exploiting logs of the user's interaction with his/her photos (e.g. changes in privacy settings, removal of previously shared images, etc.).

Provided that sufficient amount of feedback is available from each user, one could rely only on user-specific examples for training personalized privacy classification models. This, however, might require considerable effort from the user and cannot be taken for granted. As a result, user-specific privacy classification models might not be able to generalize well. To overcome this problem, we propose the development of semi-personalized models that are learned using a combination of user-specific training examples and examples from other users. The intuition behind such an expansion of the training set is that, although each person has a personal notion of privacy, there are also similarities between different users (since everyone is affected to some degree by general trends and norms) and the expansion of the training set is tailored exactly towards an exploitation of these similarities. Importantly, in order to retain the personalized nature of the models, we assign higher weights to the user-specific examples, effectively increasing their influence on the resulting model.

More formally, given a set of users $U = \{u_1, u_2, \ldots, u_k\}$ and assuming that each user $u_i \in U$ has provided ground truth annotations for a set of personal images $I_{u_i} = \{im^1_{u_i}, im^2_{u_i}, \ldots, im^n_{u_i}\}$, a user-specific dataset $D_{u_i} = \{(\mathbf{x}^1_{u_i}, y^1_{u_i}), (\mathbf{x}^2_{u_i}, y^2_{u_i}), \ldots, (\mathbf{x}^n_{u_i}, y^n_{u_i})\}$ can be constructed where $\mathbf{x}_{u_i} = [x_{1_{u_i}}, x_{2_{u_i}}, \ldots, x_{d_{u_i}}]$ is a vector representation of im_{u_i} and y_{u_i} equals 1 if the image is annotated as private, 0 otherwise. The typical approach is to train a personalized classifier $h_{u_i} : \mathcal{X} \rightarrow \mathcal{Y}$ (where $\mathcal{X} = R^d$ and $\mathcal{Y} = \{0, 1\}$ are the domains of $\mathbf{x}$ and y respectively) using only examples from D_{u_i}. Instead of that, we propose that each classifier h_{u_i} is trained on $\bigcup_{i=1}^{k} D_{u_i}$, i.e. the union of all user-specific datasets, and personalization is achieved by assigning a higher weight w to the examples of D_{u_i}. Example weights are directly handled by some learning algorithms (e.g. decision trees) while other learning algorithms can be "forced" to take weights into account by including duplicates of specific examples in the training set. The effect of weighting is that the classifier is biased towards correct prediction of higher weighted examples and is commonly used in supervised learning techniques, e.g. cost-sensitive learning [6] and boosting [8].

We note that our approach resembles techniques from the domains of *transfer* and *multi-task* learning [17, 5], commonly referred to as *instance sharing* or *instance pooling*. In fact, if we consider the privacy classification of the images of each user as a different learning task, the problem of personalized image privacy classification can be considered as an instance of multi-task learning. These methods are known to work better than methods that treat each learning task independently whenever the tasks are related and there is lack of training data for some of the tasks [1], two conditions that hold in the problem that we tackle here.

3.2 A Realistic Image Privacy Benchmark

The PicAlert dataset is certainly useful for training models that capture a generic notion of privacy. However, there are two limitations that make PicAlert unsuitable as a realistic image privacy classification benchmark: a) it consists of publicly available images with few of them being of really private nature, b) the ground truth collection process makes the unrealistic assumption that all OSN users have common privacy preferences. As a result, a privacy classification model trained on this dataset may practically fail to provide accurate classifications (as shown in Section 4.2). Moreover, the variability of privacy preferences among users is not taken into account when evaluating the accuracy of

[7]"Private are photos which have to do with the private sphere (like self portraits, family, friends, your home) or contain objects that you would not share with the entire world (like a private email). The rest is public." [24]

privacy classifications on PicAlert, resulting to overly optimistic performance estimates.

To overcome these limitations, we created a new privacy-oriented image dataset with two goals: a) the development of personalized image privacy models, and b) the realistic evaluation of both generic and personalized image privacy models. To this end, we conducted a realistic user study where we asked users to provide privacy annotations for photos of their personal collections. A call for contributions that described our research goals and the potential benefits for OSN users was distributed within our workplaces and through our OSN accounts. To reduce the concerns associated with sharing personal images (especially private ones), we provided users with software that automatically extracts the above visual features from the images and helps them share the features and the corresponding annotations (instead of the original images). To provide loose guidance and let users develop their own notion of privacy, we briefly described as public "images that they would share with *all* their OSN friends or even make them publicly visible" and as private "images that they would share only with *close* OSN friends or not share them at all". To ensure a representation of both classes we asked each user to provide (if possible) at least 10 private and 30 public images.

In total, we received feedback from 27 users (22 males and 5 females), with ages ranging from 25 to 39 years. Each user contributed approximately 16.4 private and 39.5 public photos (on average) for a total of 1511 photos. The resulting dataset (features and privacy annotations), named *YourAlert*, is made publicly available[8] for future benchmarks.

3.3 Visual and Semantic Features

In our experiments we focus on privacy classification based on the visual content - a piece of information that is always available in contrast to metadata and manually assigned tags - and extract the following state-of-the-art visual features:

`vlad`: We used the implementation of [20] to extract $d = 24,576$-dimensional VLAD+CSURF vectors from a 128-dimensional visual vocabulary and then performend PCA and whitening to project the vectors to $d' = 512$ dimensions (a projection size that led to near optimal results in preliminary experiments).

`cnn`: standard convolutional neural network features using the VGG-16 model [19] that includes 16 layers and is learned with the training set of the ImageNet ILSVRC 2014 dataset [18]. VGG-16 was chosen because it obtained one of the top results during the ImageNet 2014 challenge but also because it is publicly available and thus facilitates reproducibility. This dataset includes 1,000 specific classes and approximately 1,2 million images. These classes cover a wide range of domains and the obtained model has thus good performance in transfer learning tasks as attested by [9]. We use the output of the last fully connected layer (*fc7*), which consists of 4,096 dimensions.

`semfeat`: semantic image features obtained by exploiting the outputs of a large array of classifiers, learned with low-level features [2]. We use the VGG-16 features described above as basic features for the semantic features. Here, we compute a slightly modified version of the `semfeat` descriptor that was introduced in [9]. Only concepts that have at least 100 associated images are retained and the total size of the descriptor is 17,462. Concept models are learned independently as binary classifiers but with a ratio of 1:100 between positive and negative examples instead of a fixed number of negatives. The negative class includes images that illustrate ImageNet concepts that were not modeled. These images are sorted in order to provide a conceptually diversified sample of negatives for each modeled concept. Following the conclusions of [9] concerning the positive effect of sparsity, only the top $n = 100$ classifier outputs are retained for each image.

[8]https://github.com/MKLab-ITI/image-privacy

Table 1: Privacy-related latent topics along with the top-5 `semfeat` concepts assigned to each topic

Topic	Top-5 `semfeat` concepts assigned to each topic
children	dribbler child godson wimp niece
drinking	drinker drunk tippler thinker drunkard
erotic	slattern erotic cover-girl maillot back
relatives	g-aunt s-cousin grandfather mother g-grandchild
vacations	seaside vacationer surf-casting casting sandbank
wedding	groom bride celebrant wedding costume

Table 2: Dataset statistics

Dataset	# examples (private/public)	Source
PicAlert	26458 (3651/22807)	[24]
YourAlert	1511 (444/1067)	This paper

Compared to `vlad` and `cnn`, `semfeat` have the advantage that they enable result explainability: users can obtain human-understandable feedback about why an image was classified as private or not, in the form of top concepts associated to it. A limitation of this approach is that, having been constructed for general purpose concept detection, the `semfeat` vocabulary contains many concepts that are too specific and unrelated to privacy (e.g. *osteocyte*: 'mature bone cell'). As a result, many of the top n concepts of each image can not be easily linked to privacy.

To address this limitation, we developed a privacy aspect modeling approach that projects the detected `semfeat` concepts into a number of privacy-related latent topics using Latent Dirichlet Allocation (LDA) [3]. More specifically, each image is treated as a document consisting of its top $n = 10$ `semfeat` concepts and a private image corpus is created by combining the private images of the PicAlert and YourAlert datasets. LDA (the Mallet implementation [13]) is then applied on this corpus to create a topic model with 30 topics. Among the detected topics, 6 privacy-related ones are identified: `children`, `drinking` , `erotic`, `relatives`, `vacations`, `wedding` (Table 1). Given such a topic model, the topics of each image are inferred (using Gibbs sampling inference) from its `semfeat` concepts and the assignments to the privacy-related topics are used as a means of justification of the classifier's decision (as shown in Figure 2). We refer to this representation as `semfeat-lda`.

4. EXPERIMENTS

4.1 Experimental Setup

In our experiments we used the PicAlert[9] and YourAlert datasets, of which the statistics are provided in Table 2. To measure the accuracy of a classification model, we use

[9]Since some images of PicAlert are no longer available in Flickr, the version that we use here contains about 18% less images than the original one.

the area under the ROC curve (AUC). This was preferred over other evaluation measures due to the fact that it is unaffected by class imbalance and it is independent of the threshold applied to transform the confidence (or probability) scores of a classification model into hard 1/0 (private/public) decisions. Moreover, AUC has an intuitive interpretation: it is equal to the probability that the classification model will assign a higher score to a randomly chosen private image than a randomly chosen public image. Thus, a random classifier has an expected AUC score of 0.5 while a perfect classifier has an AUC score of 1.

Throughout the experiments, we use an L2-regularized logistic regression classifier (the LibLinear implementation [7]) as it provided a good trade-off between efficiency and accuracy compared to other state-of-the-art classifiers in preliminary experiments. Moreover, the coefficients of a regularized logistic regression model are suitable for identifying features that are strongly correlated with the class variable [10], thus facilitating explanation of the privacy classifications when features with a semantic interpretation such as `semfeat` are used. The regularization parameter was tuned by applying internal 3-fold cross-validation and choosing the value (among $10^r : r \in \{-2, \ldots, 2\}$) that leads to the highest AUC. Finally, all feature vectors were normalized to unit length before being fed to the classifier, as suggested in [7].

To facilitate reproducibility of our experimental results we have created a GitHub project[10] where we make available the experimental testbed and the datasets that we used.

4.2 Limitations of Generic Privacy Models

In this section, we evaluate the performance of generic image privacy classification models when applied in a realistic setting where different users have different perceptions of image privacy. To this end, we conduct the following experiment: A generic privacy classification model is trained using a randomly chosen 60% of the PicAlert dataset and then tested on: a) the remaining 40% of PicAlert and b) the YourAlert dataset. In the first case, we have an idealized evaluation setting (similar to the one adopted in [24]), while in the second case we have an evaluation setting that better resembles the test conditions that a privacy classification model will encounter in practice. To ensure reliability of the performance estimates, we repeat the above evaluation procedure five times (using different random splits of PicAlert) and take the average of the individual estimates.

Figure 3 shows the AUC scores obtained on PicAlert (light blue bars) and YourAlert (orange bars) when each of the visual features described in Section 3.3 is used. On PicAlert, we also evaluate the performance with quantized SIFT (`bow`) and edge-direction coherence (`edch`) features, the best performing of the visual features used in [24][11].

The performance on PicAlert indicates that `vlad`, `semfeat` and `cnn` lead to significantly better results than `edch` and `bow`. With `semfeat` and `cnn`, in particular, we obtain a near-perfect 0.95 AUC score which is about 20% better than the AUC score obtained with `bow` (the best visual feature among those used in [24]). `semfeat` have very similar performance with `cnn`, a fact that makes them a very appealing choice, given their sparsity and interpretability properties.

[10]https://github.com/MKLab-ITI/image-privacy

[11]`edch` and `bow` were kindly provided by the authors of [24].

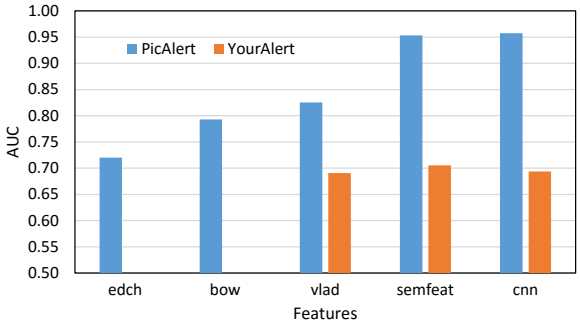

Figure 3: Performance of generic models on PicAlert and YourAlert[13]

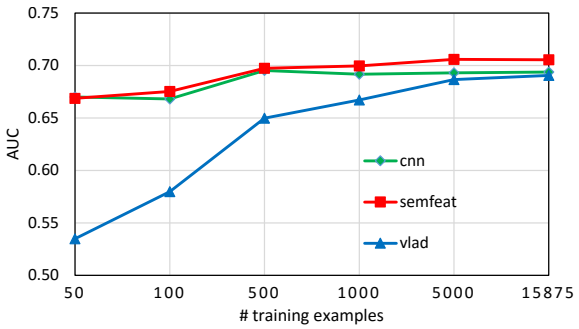

Figure 4: Performance of generic models on YourAlert as a function of the number of training examples

However, the performance of all models drops significantly when they are applied on YourAlert. `cnn` and `semfeat`, for instance, have about 24% lower performance in terms of AUC. As we see, all models perform similarly, which suggests that the accuracy is not expected to improve considerably if better features are used. Moreover, to check whether the performance on YourAlert could improve by using additional training examples from PicAlert, we studied the performance of the generic privacy models as a function of the number of examples. More specifically, for each type of features, we built six generic privacy estimation models using $\{50, 100, 500, 1000, 5000, 15875\}$ training examples from PicAlert and applied them on YourAlert. As above, to ensure reliability of the performance estimates, the evaluation of each model was repeated five times, i.e. five models were built (each trained on a different random subset of PicAlert) for each combination of features and number of training examples, and the averages of the individual performance estimates were taken. The results of this experiment are shown in Figure 4. We observe that for all types of features, the AUC performance reaches a plateau and does not change significantly after 5000 examples. Interestingly, the generic models that use `cnn` and `semfeat` features obtain 96% of their maximum performance with only 50 training examples, while the generic model that uses `vlad` features seems to require about 5000 training examples in order to approach its maximum performance. Clearly, the use of additional generic training examples is not expected to help in attaining better performance on YourAlert.

Figure 5 presents a per-user performance breakdown for generic models based on `vlad`, `semfeat` and `cnn` features (i.e.

[13]`edch` and `bow` could not be tested on YourAlert because we did not have access to their exact implementations.

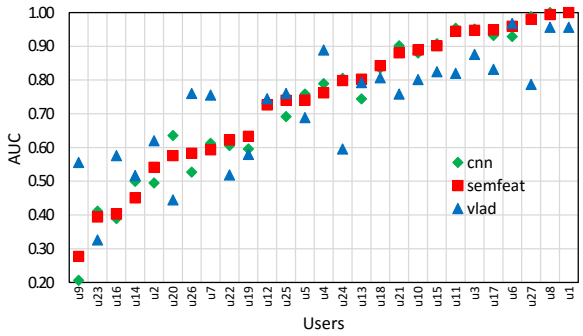

Figure 5: Per-user performance of generic models based on `vlad`, `semfeat` and `cnn` features

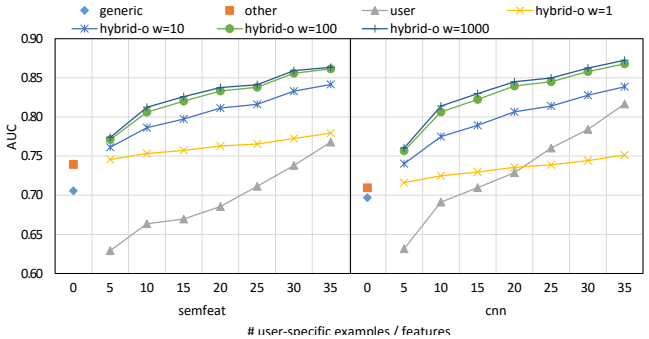

Figure 6: Performance of personalized models as a function of user-specific training examples

a separate AUC score is calculated for each user based on his/her own images). We note that there is a large variability in performance across users. For instance, using `semfeat` features, near-perfect AUC scores are obtained for users $\{u_1, u_8, u_{27}\}$ while the AUC scores are worse than random for users $\{u_9, u_{23}, u_{16}, u_{14}\}$ suggesting that the privacy perceptions of these users deviate strongly from the average notion of privacy. For this type of users, as well as for those for whom the performance of the generic models is close to random (about 40% of users), building personalized privacy classification models is essential to develop a useful alerting mechanism.

4.3 Personalized Privacy Models

This subsection compares the performance of generic privacy classification models to that of models employing user feedback in order to adapt to specific users. Specifically, we evaluate two types of personalized models on YourAlert.

`user`: Purely personalized models that use only user-specific training examples, i.e. a specific model is built for each YourAlert user from his examples only.

`hybrid`: Semi-personalized models that use a mixture of user-specific and generic training examples, with user-specific examples being assigned a higher weight to achieve personalization. We experimented with treating as generic examples: a) examples from PicAlert (`hybrid-g` variant) and b) examples from YourAlert that belong to other users (`hybrid-o` variant). Since the two choices lead to similar results, we report results only for `hybrid-o`.

As discussed in Subsection 3.1, *user* models are expected to perform better when a sufficient amount of user-specific examples are available, while *hybrid-o* models are expected to be advantageous with a limited amount of user feedback.

In order to evaluate this type of models ensuring reliable, out-of-sample estimates for all examples of each user, we use a modified k-fold cross-validation procedure ($k = 10$) that works as follows. The examples contributed by each user are randomly partitioned into k folds of approximately equal size, respecting the original class distribution (as in stratified cross-validation). Out of these, a single fold is retained as the test set and used to test the model, and from the remaining $k - 1$ folds we randomly select a specified number of examples (again respecting the original class distribution) and use them as training data either alone (`user` models) or together with generic examples (`hybrid-o` models). This process is repeated k times, with each of the k subsets being used exactly once as the test set. All predictions concerning each user are then aggregated into a single bag to calculate a per-user AUC score, or predictions for all users are combined together to calculate an overall AUC score for the examples of the YourAlert dataset.

Figure 6 plots the AUC scores on YourAlert by `user` and `hybrid-o` models trained on $\{5, 10, 15, 20, 25, 30, 35\}$ user-specific examples using `semfeat` and `cnn` features. We evaluate four variations of `hybrid-o` models, each one using a different weight ($w = \{1, 10, 100, 1000\}$) for the user-specific examples to facilitate a study of the impact of the weight parameter. In addition to the performance of these personalized models, the figure also shows the performance of two types of generic models to allow a direct comparison: a) `generic`: a model trained on a random subset of PicAlert (containing 5000 examples) and b) `other`: a model trained using only examples from other YourAlert users, i.e. a different generic model is build for each user, using the same generic examples as the corresponding `hybrid-o` model.

With respect to the generic models, we see that the performance of `other` is similar to that of `generic` with `cnn` features and better with `semfeat` features. These results suggest that although the examples of YourAlert come from users that adopt a personal, potentially different, notion of privacy, they are equally useful as the PicAlert examples for learning a generic privacy model.

With respect to the personalized models, we see that the performance of `user` models increases sharply as more user-specific training examples become available. When `semfeat` features are used we see that `user` models obtain similar performance with the generic models (`generic` and `other`) with as few as about 30 examples. The situation is even better when `cnn` features are used as we see that the performance of `user` models catches up with the performance of the generic models with as few as 15 examples and improves it by about 15% when 35 user-specific examples are used.

With regard to the semi-personalized, `hybrid-o` models we observe that they outperform significantly the purely personalized `user` models (with both types of features), especially for smaller numbers of user-specific training examples. As expected by the analysis of Subsection 3.1, the gap closes as more user-specific training examples become available. However, we see that for all values of user-specific examples (up to at least 35) `hybrid-o` models provide significantly better performance than both `user` and generic models. Importantly, we see that assigning a higher weight to user-specific examples is crucial for obtaining better per-

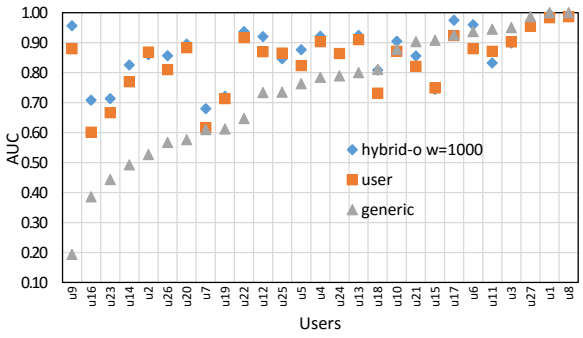

Figure 7: Per-user performance of `generic`, `user` and `hybrid-o` ($w = 1000$) models based on `cnn` features

formance. Specifically, results improve significantly as we increase w up to 100 but a saturation is observed with higher values.

Overall, the best personalized model (`hybrid-o` with `cnn` features) boosts the performance of the best generic model (`other` with `semfeat` features) by about 4% when the user provides feedback for 5 images to about 18% when the feedback increases to 35 images. Figure 7 presents a per-user performance breakdown for `generic`, `user` and `hybrid-o` ($w = 1000$) models based on `cnn` features (`user` and `hybrid-o` use 35 user-specific examples). `hybrid-o` and `user` provide better performance than `generic` for the majority of users, particularly for those that are poorly predicted by the `generic` model. Moreover, we see that in most cases `hybrid-o` is equally good or better than `user`.

4.4 Image Privacy Insights via SemFeat

Besides facilitating easily comprehensible explanations of privacy classifications (as shown in Figure 2), `semfeat` features can help in identifying users whose privacy concerns deviate strongly from the average perception of privacy. To this end, we build a single generic privacy detection model using the whole YourAlert dataset as well as 27 personalized privacy models trained using only the examples contributed by each user. For each model, we identify the concepts that are assigned the top-50 positive (associated with the private class) and top-50 negative (associated with the public class) coefficients and search for concepts that are strongly correlated to privacy according to the generic model and negatively correlated to privacy according to a personalized model (or vise versa). Despite the fact that less than 1% of the total `semfeat` features are considered in these comparisons, we can still gain valuable insights. For instance, according to the generic model, concepts related to family and relatives, such as `child`, `mate` and `son` are highly correlated to private images, while concepts related to natural scenes, such as `uphill`, `waterside` and `lakefront` are correlated to public images. In addition, we found some interesting deviations from the generic model, e.g. `alcoholic` is strongly correlated with privacy according to the generic model while it is negatively correlated with privacy for users u_{12} and u_{22}. On the other hand, concept `tourist` is private for user u_{11} and public according to the generic model.

Another practical use of `semfeat` is in creating user privacy profiles. To this end we employ the `semfeat-lda` representation that was described in subsection 3.3 and construct a privacy profile for each user by computing the centroid of the `semfeat-lda` vectors of his/her private images. This vector facilitates a summary of the user's concerns with respect to the six privacy-related topics that were identified by the LDA analysis. Given such a representation for each user, cluster analysis can be performed to identify recurring privacy themes among users. To illustrate this, we performed k-means ($k = 5$) clustering on the users of YourAlert and present the clustering results in Figure 8. We see that each cluster captures a different privacy theme. Users clustered at `c0`, for instance, are primarily concerned about preserving the privacy of their `vacations` while users clustered at `c2` are mainly concerned about the privacy of `children` and of photos related to `drinking`.

5. CONCLUSION AND FUTURE WORK

We presented a framework for personalized privacy-aware image classification. Our main immediate contribution is the creation of personalized privacy classification models that, as verified by experiments on a newly introduced image privacy dataset, exhibit significantly better performance than generic ones. Experimenting with different strategies of utilizing user feedback we found that combining user-specific with generic examples during training yields better performance than relying on either the user-specific or the generic examples alone. Furthermore, we exploited a new type of semantic features that, in addition to having an impressive performance, allow the discovery of interesting insights regarding the privacy perceptions of individuals.

There are several interesting directions for future work. First, the current system is limited to binary classification of images. We would like to develop models able to classify a user's photos into finer-grained privacy classes (e.g. close-friends, all-friends, friends-of-friends, public), corresponding to the different OSN audiences photos can be shared with. Second, we intend to design more sophisticated instance sharing strategies (e.g. assigning different weights to the examples of other users based on inter-user similarities) and make a comparison with well-established methods from the multi-task learning domain. Third, a limitation of the current study is that `cnn` features are obtained with the standard $1,000$ classes from the ImageNet challenge that are, in a large majority of cases, not linked to privacy. As an alternative, we will explore: 1) the direct training of a neural network with private/public examples, and 2) the training of a network with an increased number of privacy-oriented concepts, based on an analysis similar to the one of subsection 4.4. Such a privacy-oriented set of concepts can also facilitate more meaningful semantic justifications compared to the `semfeat` vocabulary.

In a larger context, we will work toward the structuring of an active research community working on users' privacy preservation. This effort is necessary because the tackling of the challenges presented in the introduction is only possible through the collaboration of a hefty number of research groups with expertise in different areas related to privacy. While important among the data shared in OSNs, images are just a piece of the puzzle. The mining of other types of data (texts, videos, audio content, cookies, etc.) should be combined with social network analysis in order to best serve users. As a first step, we release our implementation under an open source license and also share the features and annotations associated with our dataset to encourage reproducibility. Sharing the dataset itself would be very useful

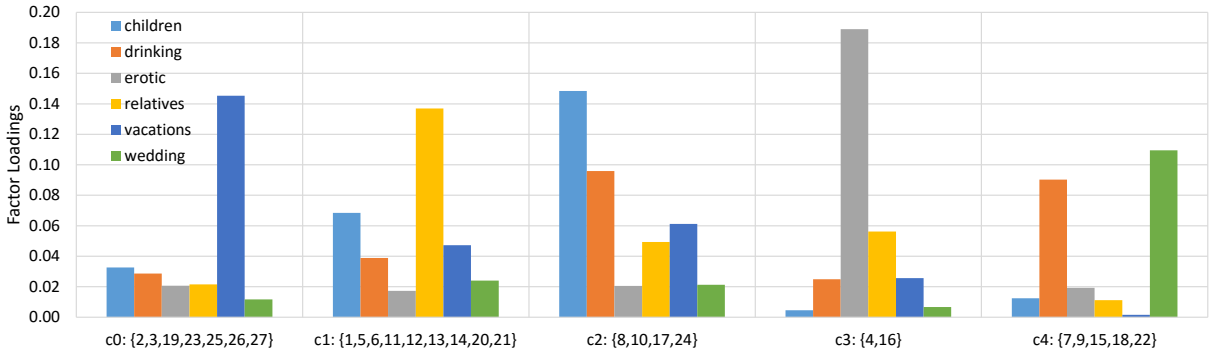

Figure 8: Clustering of YourAlert users based on privacy-related latent topics

but it is challenging due to the highly private nature of its content. The creation of dataset was not straightforward and we are currently investigating ways to enlarge it. One possible way is to provide incentives to users in exchange for the possibility to include their images in the dataset and to share them with the community. Another possibility is to liaise with other research groups from the area in order to gather data in a collaborative manner. The insights gained from this exploration will be shared with the research community in order to facilitate the creation and sharing of datasets for other types of data.

Yet another direction that we will explore is the organization of events to disseminate the topic in the community. Toward this direction, we have already identified the *MediaEval Benchmarking Initiative*[14] as a relevant venue that could host a task on privacy classification for multimedia documents.

Beyond computer science, privacy research should be informed by results from the legal and social sciences domains. A small interdisciplinary European group is already constituted as part of the USEMP project. We will work toward its extension with relevant research groups from other countries in order to include and confront different takes at privacy.

6. ACKNOWLEDGMENT

This work is supported by the USEMP FP7 project, partially funded by the EC under contract number 611596.

7. REFERENCES

[1] M. A. Alvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012.

[2] A. Bergamo and L. Torresani. Meta-class features for large-scale object categorization on a budget. In *CVPR*, 2012.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[4] D. Buschek, M. Bader, E. von Zezschwitz, and A. D. Luca. Automatic privacy classification of personal photos. In *Human-Computer Interaction - INTERACT*, 2015.

[5] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[6] C. Elkan. The foundations of cost-sensitive learning. In *IJCAI*, 2001.

[7] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[14] http://www.multimediaeval.org

[8] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *ICML*, 1996.

[9] A. Gînsca, A. Popescu, H. L. Borgne, N. Ballas, P. Vo, and I. Kanellos. Large-scale image mining with flickr groups. In *MultiMedia Modeling Conf.*, 2015.

[10] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani. *The elements of statistical learning*. Springer, 2009.

[11] Y. Liu, P. K. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing facebook privacy settings: user expectations vs. reality. In *Proceedings of the 11th ACM SIGCOMM Internet Measurement Conference*, 2011.

[12] M. Madejski, M. L. Johnson, and S. M. Bellovin. A study of privacy settings errors in an online social network. In *Tenth Annual IEEE International Conference on Pervasive Computing and Communications*, 2012.

[13] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[14] K. D. Naini, I. S. Altingovde, R. Kawase, E. Herder, and C. Niederée. Analyzing and predicting privacy settings in the social web. In *User Modeling, Adaptation and Personalization*, 2015.

[15] P. A. Norberg, D. R. Horne, and D. A. Horne. The Privacy Paradox: Personal Information Disclosure Intentions versus Behaviors. *J. of Consumer Affairs*, 41(1):100–126, 2007.

[16] C. Paine, U. Reips, S. Stieger, A. N. Joinson, and T. Buchanan. Internet users' perceptions of 'privacy concerns' and 'privacy actions'. *Int. J. Hum.-Comput. Stud.*, 65(6):526–536, 2007.

[17] L. Y. Pratt. Discriminability-based transfer between neural networks. In *NIPS*, 1992.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014.

[20] E. Spyromitros-Xioufis, S. Papadopoulos, I. Kompatsiaris, G. Tsoumakas, and I. Vlahavas. A comprehensive study over vlad and product quantization in large-scale image retrieval. *IEEE Transactions on Multimedia*, 2014.

[21] A. C. Squicciarini, C. Caragea, and R. Balakavi. Analyzing images' privacy for the modern web. In *In ACM Hypertext*, pages 136–147, 2014.

[22] A. Tonge and C. Caragea. Image privacy prediction using deep features. In *AAAI Conference on Artificial Intelligence*, 2016.

[23] S. Zerr, S. Siersdorfer, and J. S. Hare. Picalert!: a system for privacy-aware image classification and retrieval. In *ACM CIKM*, pages 2710–2712, 2012.

[24] S. Zerr, S. Siersdorfer, J. S. Hare, and E. Demidova. Privacy-aware image classification and search. In *ACM SIGIR*, 2012.