Explain to Me: Towards Understanding Privacy Decisions

Gönül Aycı
Bogazici University
Turkey
gonul.ayci@boun.edu.tr

Pınar Yolum
Utrecht University
The Netherlands
p.yolum@uu.nl

Arzucan Özgür
Bogazici University
Turkey
arzucan.ozgur@boun.edu.tr

Murat Şensoy
Ozyegin University
Turkey
drmuratsensoy@gmail.com

arXiv:2301.02079v1 [cs.AI] 5 Jan 2023

ABSTRACT

Privacy assistants help users manage their privacy online. Their tasks could vary from detecting privacy violations to recommending sharing actions for content that the user intends to share. Recent work on these tasks are promising and show that privacy assistants can successfully tackle them. However, for such privacy assistants to be employed by users, it is important that these assistants can explain their decisions to users. Accordingly, this paper develops a methodology to create explanations of privacy. The methodology is based on identifying important topics in a domain of interest, providing explanation schemes for decisions, and generating them automatically. We apply our proposed methodology on a real-world privacy data set, which contains images labeled as private or public to explain the labels. We evaluate our approach on a user study that depicts what factors are influential for users to find explanations useful.

KEYWORDS

Privacy, explainability, online social networks

1 INTRODUCTION

Managing privacy online is becoming more and more challenging. On one hand, people use systems, such as online social networks or Internet of Things applications heavily as these systems provide useful services. For example, it is common to share a document with co-authors over a Cloud service or make use of home entertainment systems that communicate with each other. On the other hand, people are worried about their privacy and think twice before using these systems. It is common for people to delete content after sharing or self-censoring themselves [6]. The problem is getting more difficult to handle as people are constantly in a situation to decide whether they would be willing to share a piece of content or not. Since the amount of content is high, people easily make errors in their decisions. Even worse, due to decision fatigue, people do not spend the time to make an informed decision. Various recent surveys conducted with users of online social networks indicate that people do not even read the privacy policies that they accept [1, 26].

Privacy assistants that work side by side with humans in a decentralized manner could serve to address this problem. Privacy assistants have been developed for various privacy assistance including checking for privacy violations [11], resolving privacy conflicts among humans [20, 25], recommending sharing policies [8, 23], and signaling if a piece of content is private [12, 24]. While doing these tasks, it is important for the privacy assistant to be able to explain its decisions to the user.

This paper considers a personal assistant that helps its user decide if a given image is private or not and proposes a methodology and a system to explain this decision to the user. One of the important works of explainability in conjunction with privacy is by Mosca and Such [20], where they develop an agent that uses computational argumentation to resolve disputes and propose a text-based description of the outcome generated by the system. However, to the best of our knowledge, there does not exist any methodology to generate explanations as to why a given content is private or public.

Existing work on explanation for binary classifications generally consider what features of the classification have been influential for the classification. Using these saliency methods, for example, heat maps can be generated such that parts of an image are highlighted to demonstrate its effect on the decision [5]. Lundberg *et al.* [17] propose a model-agnostic feature relevance explanation model, SHAP (SHapley Additive exPlanations), that is based on a game theoretically Shapley values [22]. This method computes the contribution of each feature to the prediction output. Lundberg *et al.* [16] also propose TreeExplainer that explains predictions of tree-based machine learning models. TreeExplainer is a variant of SHAP, which provides the computation of local explanations based on Shapley values in polynomial time. These approaches are important because they provide interpretability for the underlying classifier. However, they are not meant to provide explanations to the end user as we aim here.

In order to address this problem, we propose a new representation for explaining why an image is considered private or public. Our representation is made up of visually exhibiting one or more topics that the image is associated with while emphasizing important keywords that put the image in a given topic. A natural language description accompanies the visuals to describe the relation between the topics. We provide a methodology to derive these explanations from a dataset where images are labeled as private or public. We implement our methodology and apply it to a well-known image dataset for privacy. We then perform a user study to measure if users actually find these explanations useful and what factors of the explanation or the image affect users' understanding of the decision.

The rest of this paper is organized as follows. Section 2 explains our understanding of explanation, its formalization, and its relation to topic modelling. Section 3 develops our methodology into a system that can be readily used to explain privacy labels of images and evaluates the effectiveness of the extracted topics. Section 4 presents how to generate explanations from topics. Section 5 evaluates our system through a user study. Section 6 discusses our work in relation to related work. Finally, Section 7 concludes our work and provides future directions.

2 METHODOLOGY FOR EXPLAINING PRIVACY

Given an image that is classified as private or public, we would like to generate an explanation as to why this is so.

2.1 Understanding Explanation

The explanations that we are interested in generating are meant for end users. Hence, even if our explanations are influenced by the features that are used for classification, our aim is not to educate the user about how the underlying classifier works. Hence, the explanation should not be too technical. At the same time, given that many users do not read long texts on privacy policies for example, we would like the explanation to be visually understandable and supported by a short text.

Based on these constraints, we propose to formulate an explanation as to whether an image is private or public by a set of *topics* that the image belongs to. These topics are shown as a circle and labeled by the topic name. Each image can have one or more topics. Additionally, we identify one or more keywords that link this image to each topic and denote them in the corresponding topic circle. The intended understanding of this representation is that the image is private or public, because it can be described with these topics and keywords. This visual representation is augmented with a short description that falls into a predetermined language structure to explain the visual representation. The text is thus supplementary and does not provide addition information. Figure 1 shows an example image, which is annotated as public by annotators. Figure 2 shows the explanation for the image in the proposed explanation schema. The explanation provides information that the image is classified as public because it is associated with topic *Business* with the specific keyword "sign".

Figure 1: Example image annotated as public

Figure 2: The explanation for the image in Figure 1

2.2 Understanding Topics

In order to realize the above explanations, we need to understand how we can associate images with topics. Machine Learning algorithms are mostly black-box models and use a large number of features while making predictions. Thus, the models are not straightforwardly understandable for humans and are not able to make explainable predictions. Motivated by this observation, our aim is to understand the model and its predictions and develop a methodology to generate explanations for privacy decisions. Thus, for a prediction of a single instance, we need to extract the most important and relevant features from all the features in the decision. For this purpose, we propose to uncover groups of keywords (i.e., latent topics) from a collection of textual information that best represents the information in the collection. A topic consists of relevant descriptive keywords. Each image is associated with topics based on its keywords.

Topics should be meaningful and interpretable for humans. One way of realizing this during computation of topics is to ensure that the topics are *coherent*. Each topic should pertain to images that could be described with similar keywords. At the same time, each topic is relatively different from each other. We can measure coherence based on two different criteria as follows:

(i) Intra-topic similarity: The average semantic similarity between all pairs of the most associated N keywords in the same topic.
(ii) Inter-topic similarity: The average semantic similarity of the most associated N keywords from different topics.

That is, we can calculate how close the keywords that describe a topic are semantically using intra-topic similarity and how far the topics are semantically apart using inter-topic similarity. For a good topic modelling, we would want to maximize the intra-topic similarity and minimize inter-topic similarity.

3 GENERATING TOPICS

Topic Modelling is a technique that discovers latent topics within a collection of textual information. It allows us to extract different topics (features) from keyword sets.

3.1 Topic Modelling

We use a widely used topic modelling technique, namely, Non-negative Matrix Factorization (NMF) [14]. NMF is an approximation to factorize a non-negative matrix of a non-negative image-keyword matrix X, into non-negative matrices W and H as in Figure 3. W

(features) matrix stores how much each image belongs to a topic and H (components) matrix stores how much each keyword belongs to each topic. The W and H matrices are initialized randomly. NMF algorithm runs iteratively until it finds a W and H that minimize the Frobenius norm of the matrix, that is, $\|X - W \times H\|_F$. NMF is suitable for interpretability (components are non-negative) and works better and faster for short texts (a set of keywords) as compared to alternatives such as Latent Dirichlet allocation (LDA) [4]. In this study, we make use of the term weighting method, namely, the Term Frequency - Inverse Document Frequency (TF-IDF) model to transform keywords into numerical vectors in order to construct an *image-keyword (X)* matrix. TF-IDF assigns weights based on how relevant a keyword is to a given collection of keyword sets. We build the NMF model for a different number of topic (k) values, which generates an *image-topic (W)* matrix and a *topic-keyword (H)* matrix, and then we use the Random Forest algorithm to make predictions.

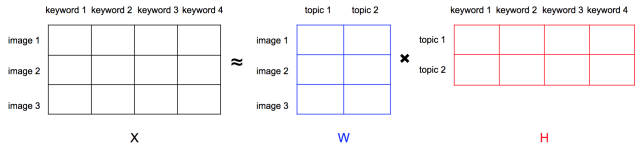


Figure 3: Non-negative Matrix Factorization (NMF) concept

3.2 Evaluation of Topics

We use a balanced subset of the publicly available PicAlert dataset [27]. The PicAlert is a well-known and widely used dataset for the privacy prediction problem for images [3, 12, 23, 24]. It contains Flickr images that are labeled as *private* or *public* by annotators. These images are labeled by 81 users between 10 and 59 years of age with different backgrounds. We consider an image as private if at least one annotator has annotated it as private and public if all the annotators have annotated it as public. The balanced subset we work with contains 32*K* samples, including 27*K Train* and 5*K Test*, which are labeled as *private* or *public*. Then, we automatically generate 20 different descriptive keywords for each image using Clarifai API [1].

In the NMF model, we set the number of topics based on the model performance in terms of *coherence*. While calculating intra-topic similarity and inter-topic similarity, each keyword is represented by word embedding vectors, namely, word2vec [18]. The similarity between two keyword vectors is measured by the Cosine-Similarity metric. Semantically similar tags tend to be close to each other in the semantic space. Intra-topic similarity values for 20 topics and 10 topics are 0.20 and 0.18, respectively. Additionally, inter-topic similarity values for 20 topics and 10 topics are 0.43 and 0.48. This indicates that topics are more distinguishable from other topics for $k = 20$ as compared to $k = 10$, while also keywords in the topics are closer to each other. We represent keywords as 300−dimensional vectors of the word2vec model trained on Google News when calculating coherence. Note that the cosine-similarity values between two vectors for this model are generally low (e.g., the similarity between "person" and "people" is 0.51 and "tree" and "park" is 0.23).

We named 20 topics that we discover using NMF. Figure 4 shows keyword clouds for five different topics (i.e., *Nature, Child, Performance, Business,* and *Fashion*) with the top 20 keywords that describe each topic. The font size is sensitive to relative significance. That is, the most descriptive keyword is displayed as the largest. For instance, the top five descriptive keywords of the topic *Nature* are $\{tree, park, wood, nature, outdoors\}$. Figure 5 shows the percentage of each topic being associated with private and public images. Some topics such as the topic *People* are associated more frequently with the private class, whereas some of them such as *Sky* are associated more frequently with the public class. Note that although some topics are associated more frequently with private and some with public images, the topics do not have an explicit class to which they belong. Therefore, the topic itself does not directly signal a certain class and thus, it is not straightforward to generate an explanation for the decision only by looking at its class.

To evaluate the representation of the images with the topics extracted using NMF, we trained a Random Forest classifier where the images are represented as TF-IDF vectors of these topics. The classifier yields an accuracy of 88.5% on the test set, indicating that the NMF-extracted topics are effective for privacy prediction. It achieves state-of-the-art performance for image privacy prediction compared to existing approaches [3, 24].

4 GENERATING EXPLANATIONS FROM TOPICS

The TreeExplainer [16] model provides the contributions of each feature in terms of Shapley values, which affect the model output of tree-based algorithms such as Random Forest. Not all features have equal contribution to a class prediction: a feature can push the prediction higher (positive Shapley value) or lower (negative Shapley value). The machine learning model concludes its prediction by taking into account the contribution of each feature. This is useful in interpreting how the classifier works. One way to create explanations would be to display all these values to the user. However, as the number of features increases, it would be cumbersome and confusing to show them all to the end user. Therefore, we start from the TreeExplainer idea, but modify it to match our expectations for explanations, as described in Section 2.

In this study, each feature corresponds to a topic. We are interested in identifying topics that are useful in explaining the content of the image at hand. For example, for a given image, a large positive Shapley value might be assigned to a topic because the image is related to that topic. But, it might also be the case that a large negative value is assigned to a topic that is unrelated to the topic. The second category shows that the classifier made a decision based on the fact that the image did not exhibit the properties associated with this topic. While useful to understand the classifier, this information is difficult and possibly unnecessary to show to the user. Hence, we need to carefully decide how to use the Shapley values when creating the explanations.

Our methodology generates human-understandable explanations through topic reduction in the output of the TreeExplainer

[1] https://www.clarifai.com/

(a) Nature (b) Child (c) Performance (d) Business (e) Fashion

Figure 4: Keyword clouds for Topics Nature, Child, Performance, Business, and Fashion

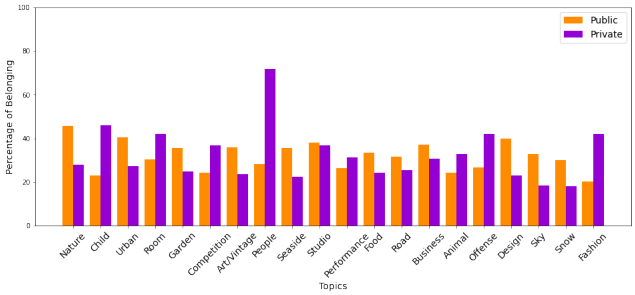


Figure 5: Percentage of occurrence of each topic in private and public images

model. Regarding topic reduction, we divide images into four categories: *Dominant*, *Conflicting*, *Collaborative*, and *Vague*, in terms of the contribution of topics to the decision.

Dominant: An image belongs to the *Dominant* category when the contribution of one topic is decisive for the class prediction. That is, a topic makes a relatively high contribution compared to other topics of the image.

Figure 6 shows an example image in the Dominant category that has been identified as private by annotators. The generated explanation for this image being assigned to the private class is that it is relevant to the topic Child with the keywords including $\{child, fun, son, happiness, family\}$.

Algorithm 1: Find Dominant Topics

Input : $P \in \mathbb{N}$, the number of images
$T \in \mathbb{N}$, the number of topics
$normalized\_matrix \in \mathbb{R}^{P \times T}$, stores the normalized Shapley values of the associated topics for each image
$d\_ub \in [0, 1]$, the upper bound

Output: $idx\_dominant$, the set of the indexes of images that have Dominant Topics

```
1 idx_dominant ← {∅}
2 for p = 1 to P do
3     for t = 1 to T do
4         if normalized_matrix[p][t] ≥ d_ub then
5             idx_dominant ← idx_dominant ∪ {p}
          else
6             do nothing
```

The generated explanation for this image being assigned to the private class is that it is related to the topic **Child** with these specific keywords.

Figure 6: Example image annotated as private and its generated explanation with the topic Child (Dominant category)

Algorithm 1 describes how to find images belonging to the dominant category, that is, the images with dominant topics. In this algorithm, P and T correspond to the number of images and topics, respectively. $normalized\_matrix$ is the matrix that contains normalized Shapley values of topics for each image. d_ub is the upper bound with respect to deciding a Dominant topic. First, we initialize the output of the Algorithm 1, $idx\_dominant$, to store the indexes of images with dominant topics (line 1). We normalized the Shapley values by dividing each Shapley value by the sum of the absolute values of all topics of the image. The algorithm enlarges the $idx\_dominant$ set when the normalized Shapley values of the associated topic with the image are greater than or equal to the threshold (lines 4 – 5). For the evaluations, we set this threshold to 0.7.

Collaborative: An image belongs to the *Collaborative* category when the contributions of its topics arrive at a consensus about the decision. That is, the images in this category do not have a

single decisive topic as in Dominant, but have topics that support each other collaboratively. Figure 7 shows an example image that has been identified as private. The generated explanation by our algorithm for this image being assigned to the private class is that it is relevant to the topics People, Fashion, and Room with certain keywords shown in the topic circles. All three topics push the prediction higher. The Algorithm 1 extended by N topics is also applicable to finding such images that belong to the collaborative category. That is, the total contributions of N topics are decisive.

The generated explanation for this image being assigned to the private class is that it is related to the topics **People, Fashion**, and **Room** with these specific keywords.

Figure 7: Example image annotated as private and its generated explanation with the topics People, Fashion, and Room (Collaborative category)

Conflicting: The topics associated with an image do not always agree on whether the image should be private or public. In such situations, the explanation should indicate this. An image belongs to the *Conflicting* category if the image has topics whose magnitudes are almost equal but the contributions to a class prediction are in the opposite direction. Making a decision can be difficult when an image has conflicting topics that have opposing forces in the decision.

Figure 8 shows an example image that has been identified as public by annotators. The generated explanation for this image is that even though it is relevant to the topic People with the specific keywords (i.e., "wear", "man", "people"), it is also relevant to the topic Art/Vintage that pushes the prediction higher and for that reason, it is classified as public.

Algorithm 2 describes the process of finding images with conflicting topics. In this algorithm, P and T correspond to the number of images and topics, respectively. We normalized the Shapley values by dividing each Shapley value by the sum of the absolute values of all topics of the image. *topic_matrix* is the matrix that contains the Shapley values of topics for each image and *normalized_matrix* stores the normalization of *topic_matrix*. *c_ub* is the upper bound

The generated explanation for this is that even though it is related to the topic **People** with the specific keywords below (which signals the private class), it is also related to the topic **Art/Vintage** and for that reason, it is classified as public.

Figure 8: Example image annotated as public and its generated explanation with the topics Art/Vintage and People (Conflicting category)

with respect to deciding a Conflicting topic. First, we initialize the output of Algorithm 2, $idx\_conflict$, to store the indexes of images with conflicting topics (line 1). For each image, the algorithm tries to find topics that push predictions high and low and also magnitudes of these contribution are greater than or equal to the threshold, 0.2 (lines $1-9$). The algorithm enlarges the $idx\_conflict$ set when there exist at least two such topics in the image (lines $10-12$).

Vague: It is also possible that an image belongs to many topics with a low confidence. Thus, it would not fall into any of the above three categories. Therefore, its class cannot be explained as clearly as the others. We call this category *Vague* and generate an explanation that contains the top topics. That is, if there are topics whose contributions are relatively small, we ignore them. In doing so, our aim is to generate explanations with the most relevant and influential topics for the decision.

Figure 9 shows an example image that has been identified as private by annotators. The generated explanation for this image is that even though it is related to the topic Urban with the specific keywords (i.e. "urban", "city", "street"), it is also relevant to the topics People and Offense and for that reason, it is classified as private.

5 EVALUATION

We performed an online user study to evaluate our proposed explanation model in terms of sufficiency, satisfaction, and understanding. We conducted a pilot study (with $n = 5$ users) before the real study to test whether the study is understandable. Based on the comments during the pilot, we improved the initial description of the study and reworded one question.

Algorithm 2: Find Conflicting Topics

Input : $P \in \mathbb{N}$, the number of images
$T \in \mathbb{N}$, the number of topics
$topic\_matrix \in \mathbb{R}^{P \times T}$, stores the Shapley values of the associated topics for each image
$normalized\_matrix \in \mathbb{R}^{P \times T}$, stores the normalized Shapley values of the associated topics for each image
$c\_ub \in [0, 1]$, the upper bound

Output : $idx\_conflict$, the set of the indexes of images that have Conflicting Topics

```
idx_conflict ← {∅}
for p = 1 to P do
    positive_count ← 0
    negative_count ← 0
    for t = 1 to T do
        if normalized_matrix[p][t] ≥ c_ub then
            if topic_matrix[p][t]) > 0 then
                positive_count ← positive_count + 1
            else if topic_matrix[p][t]) < 0 then
                negative_count ← negative_count + 1
            else
                do nothing
    if positive_count > 0 and negative_count > 0 then
        idx_conflict ← idx_conflict ∪ {p}
    else
        do nothing
```

5.1 User Study

Our user study has three phases. In the first phase, we present a plain language statement that describes the study and a consent form. The second phase is meant to explains the study over an example, wherein we show an image, its generated explanation, and the three questions that will be asked to the participant. Finally, in the third phase, each participant is exposed to 16 images with generated explanations in a random order. Two of these images deliberately provide irrelevant explanations so that we can differentiate the participants that are attentive during the survey. Thus, these questions are meant to filter out the participants who are not focused. Such users are removed from the analysis.

In order to examine our explanation model, we personalize the *Explanation Satisfaction Scale* proposed by Hoffman *et al.* [10]. We ask participants to rank the following questions:

(1) This explanation that the algorithm produces has SUFFICIENT DETAIL.
(2) This explanation produced by the algorithm is SATISFYING.
(3) From this explanation, I UNDERSTAND why an image has been identified as private or public.

Each factor is accompanied by a 5−point Likert scale (Strongly agree = 5, Somewhat agree = 4, Neither agree nor disagree = 3, Somewhat disagree = 2, Strongly disagree = 1). In the final phase, participants responded to anonymously collected demographic

Figure 9: Example image annotated as private and its generated explanation with the topics People, Urban, and Offense (Vague category)

questions (age, gender, and education level) and optionally provided free-form text for comments/feedback. We designed our user study using the Qualtrics online survey tool [2].

5.2 Participants

A total of 57 participants responded to questions but we excluded 12 of them who did not catch the check questions properly. 64% of the remaining 45 participants were male and 36% were female. 26 participants were between 25-34 years old, 14 were between 18-24, 4 were between 35-44, and 1 was between 55-64. In terms of the highest degree of education, 19 of them had a Master's degree, 11 of them had Bachelor's degree, 6 of them were High school graduates, 5 of them attended Some college (1-4 years, no degree), 2 of them had Doctorate degree, and 2 of them had Professional school degree (MD, DDC, JD, etc).

5.3 Results

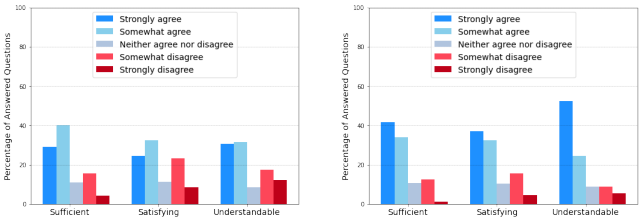
The performed user study shows that generated explanations are useful for and make sense to humans. Table 1 demonstrates how confidence levels change based on intervals of mean value. *Sufficient, Satisfying,* and *Understandable* in Figures 10 and 11 correspond to the question 1, 2, and 3 in Section 5.1, respectively. Our results indicate that participants were very confident that explanations were sufficiently detailed ($M = 3.88, SD = 1.12$), found the explanations satisfactory ($M = 3.62, SD = 1.28$), and understood why the images were labeled as private or public ($M = 3.8, SD = 1.33$). We

[2] https://www.qualtrics.com

Interval Level	Confidence Level
1.00 - 1.79	Not at all confident
1.80 - 2.59	Slightly confident
2.60 - 3.39	Moderately confident
3.40 - 4.19	Very confident
4.20 - 5.00	Extremely confident

Table 1: Confidence Levels based on Mean

evaluate the performance of our methodology based on the private and public classes and different image categories such as dominant, collaborative, conflicting, and vague.



(a) Private (b) Public

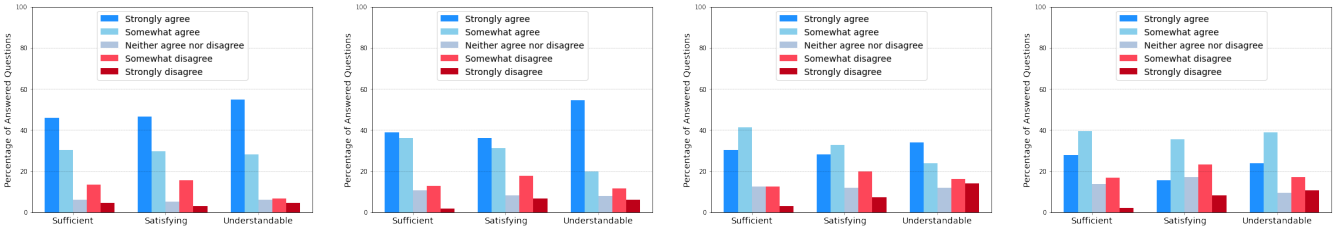
Figure 10: Distribution of answers with respect to the classes. The explanations for images that are labeled as public have been found to be more sufficient, more satisfying and more understandable compared to the images labeled as private.

Figure 10 shows the distributions of answers for the survey questions with respect to the private and public classes. Figure 10(a) indicates that participants were very confident that the generated explanations for the private class images are sufficient, satisfying, and understandable. For instance, from the explanations for private images, they understood why images have been identified as private ($M = 3.51, SD = 1.39$). However, 10(b) indicates that participants found the explanations for public images to be more sufficient, more satisfying, and more understandable compared to the images labeled as private. For instance, participants understand better why images have been identified as public ($M = 4.1, SD = 1.2$). This is inline with recent work [3], which has shown that privacy is inherently ambiguous and their personal privacy assistant yields better performance for the public class.

Figure 11 shows the distributions of answers to assess sufficiency, satisfaction, and understandability with respect to different categories (i.e., Dominant, Conflicting, Collaborative, and Vague). Figure 11(a) and 11(b) demonstrate when an explanation has a decisive topic or is composed of like-minded topics in the decision, participants are very confident that the explanations of images belonging to such categories are sufficiently detailed and satisfying. Additionally, participants are confident about understanding why an image is identified as belonging to a certain class (private or public). On the other hand, compared to the Dominant and Collaborative categories, Figure 11(c) shows that participants are less confident ($M = 3.47, SD = 1.45$) about understanding the decision when an explanation has topics that have opposing forces in the decision. The images in this category have conflicting topics in terms of the contribution to the decisions. Thus, making a decision is not straightforward for the images whose explanations belong to the Conflicting category as compared to the Dominant and Collaborative categories. Moreover, Figure 11(d) shows the results for the explanations of the images belonging to the Vague category. Even if participants are moderately confident ($M = 3.27, SD = 1.22$) that the explanations are satisfying, they are very confident about the sufficiency of the explanations and understandability of a class decision based on the explanations.

6 DISCUSSION

In the literature, several studies on image privacy prediction make use of descriptive keywords (tags) and visual features. Squicciarini *et al.* [23] present a Tag-To-Protect (T2P) system that automatically recommends privacy policies using the image tags. Their experiment shows that prediction accuracy decreases when there are large tag sets and when the number of tags per image increases. Tonge and Caragea [24] use deep visual semantic (i.e., deep tags) and textual features (i.e., user tags) to develop a model to predict the privacy of images as private or public. They use Support Vector Machine (SVM) classifiers with pre-trained CNN architectures such as AlexNet, GoogLeNet, VGG-16, and ResNet to extract features (tags). Deep tags of images are the top k predicted object categories extracted from pre-trained models. Using user-created tags, they create deep visual features by adding highly correlated tags to visual features extracted from the fully connected layer of the pre-trained models. They find that a combination of user tags and deep visual features from ResNet with the top 350 correlated tags performs the best. Kurtan and Yolum [12] propose an agent-based approach, namely PELTE, which addresses the same problem with automatically generated image tags. The internal tag table stores the data of privacy labels collected from images shared by the user itself. The external tag table stores the data of images shared by the user's friends. Their proposed system performs well in predicting privacy, even though the personal assistant only has access to a small amount of data. Ayci *et al.* [3] propose a personal privacy assistant called PURE to preserve the privacy of its user. PURE is aware of uncertainty by generating an uncertainty value for each prediction of a given image, informing its user about it, and delegating decisions back to the user if it is uncertain about its predictions. PURE is able to make personalized predictions by using the personal data of its user. It is also risk-averse, by incorporating the user's risk of misclassification. Their experiments are fruitful in analyzing the link between uncertainty and misclassification. They show that PURE captures uncertainty well and performs better compared to alternative models for quantifying uncertainty (i.e., Monte Carlo dropout [9] and Deep Ensemble [13]). Although they demonstrate the success of using descriptive keywords and visual features to predict image privacy, neither of these approaches address capturing the explanations for the privacy predictions as we have done here. However, explaining the model predictions is critical to understanding people's privacy expectations and preferences. In this study, we propose a novel methodology that uses descriptive keywords to explore latent topics by topic modelling and provides explanation schemes for predictions.



(a) Dominant (b) Collaborative (c) Conflicting (d) Vague

Figure 11: The answers for the questions with respect to the categories

Dammu *et al.* [7] develop a personalized privacy prediction system that is personalizable, explainable, configurable, and comes with customizable privacy labels. The system consists of four modules: object detection, location detection, object localization, and explicit content extraction. The decision network aggregates the modules' outputs for personalized privacy predictions. This approach enables personalized image predictions by incorporating user feedback. However, it is not yet clear how this approach can scale in applications that use large image sets. Miller [19] examine studies of explainability within the scope of philosophy, social and cognitive psychology, and cognitive science. Their study provides various definitions of explainability, criteria for selecting explanations, evaluating explanations, and useful insights for Explainable Artificial Intelligence (XAI). They define interpretability of a model as the degree to which the cause of a prediction can be understood. Explainability is defined by the interpretability that one can adopt and the understanding of the explanation obtained. Justification is provided by explaining why a decision is good. Responsibility is one of the criteria for selecting explanations, indicating what caused an event to occur and the minimum number of changes that must be made to prevent that event from occurring. Arrieta *et al.* [2] provide an overview from a broad perspective on XAI by defining interpretability and explainability. They define interpretability as the ability to explain meaning in a form that people can understand. They associate explainability with explanation as the interface between a human and a decision maker. They provide a taxonomy for explainability techniques in machine learning (ML) models. They examine XAI in ML, which captures transparent models (i.e., linear regression or Bayesian models) and post-hoc explainability techniques that can be both model-agnostic and model-specific. In general, these techniques include model simplification (e.g., rule extraction methods), feature relevance explanation, and visual explanation. While they use all features in the explanations, this is not always straightforwardly understandable. We develop a powerful methodology that is capable of generating explanations with only relevant topics.

Orekondy *et al.* [21] present a model for the privacy risk prediction task for images and provide 68 privacy attributes such as *nudity, passport* and *religion*. Li *et al.* [15] propose a method to find out what kind of visual content is private. They develop a taxonomy with 28 categories such as *nudity/sexual, irresponsible to child* and *bad characters/unlawful/criminal*. Zhao *et al* [28] define a privacy taxonomy with 10 categories with the most commonly used descriptive keywords for a certain category. For example, the descriptive keywords of the category *religion/culture* include *culture, religion and spiritual*. Even though they propose inspiring taxonomies for privacy by synthesizing existing literature, their approaches do not provide explanations for a particular image as to why the image is labeled private or public, as we have done here. Moreover, we use topic modelling to explore hidden topics that are not associated only with a particular class.

7 CONCLUSION

In this paper, we propose a novel methodology to understand why a given image is private or public. Our method is able to explore latent topics using topic modelling from descriptive keywords of images. It makes privacy predictions based on the relationship between images and their associated topics, and automatically generates explanations for privacy decisions. The privacy classifier achieves high accuracy, demonstrating the effectiveness of the topic-based representation of images. Based on a user study, we show that the generated explanations make sense to people and that participants find the explanations sufficient, satisfying, and understandable. An important direction for future work is to be able to get feedback from people and update the explanations. Another interesting direction would be to incorporate prediction uncertainty [3] into our proposed methodology. In this way, the uncertainty identified in the images can be explained to the user. Further feedback from the user could help the system decrease the uncertainty for future predictions.

REFERENCES

[1] Alessandro Acquisti and Jens Grossklags. 2005. Privacy and rationality in individual decision making. *IEEE security & privacy* 3, 1 (2005), 26–33.
[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.
[3] Gönül Aycı, Murat Şensoy, Arzucan Özgür, and Pınar Yolum. 2022. Uncertainty-Aware Personal Assistant for Making Personalized Privacy Decisions. *ACM Transactions on Internet Technology* (August 2022).
[4] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[5] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. 2019. Salient object detection: A survey. *Computational visual media* 5, 2 (2019), 117–150.
[6] Philip Cook and Conrad Heilmann. 2013. Two types of self-censorship: Public and private. *Political studies* 61, 1 (2013), 178–196.
[7] Preetam Prabhu Srikar Dammu, Srinivasa Rao Chalamala, and Ajeet Kumar Singh. 2021. Explainable and Personalized Privacy Prediction. (2021).

[8] Ricard L Fogues, Pradeep K Murukannaiah, Jose M Such, and Munindar P Singh. 2017. Sosharp: Recommending sharing policies in multiuser privacy scenarios. *IEEE Internet Computing* 21, 6 (2017), 28–36.

[9] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. PMLR, 1050–1059.

[10] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[11] Nadin Kökciyan and Pınar Yolum. 2016. P ri g uard: A semantic approach to detect privacy violations in online social networks. *IEEE Transactions on Knowledge and Data Engineering* 28, 10 (2016), 2724–2737.

[12] A Kurtan and Pınar Yolum. 2021. Assisting humans in privacy management: an agent-based approach. *Autonomous Agents and Multi-Agent Systems* 35, 1 (2021), 1–33.

[13] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* 30 (2017).

[14] Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788–791.

[15] Yifang Li, Nishant Vishwamitra, Hongxin Hu, and Kelly Caine. 2020. Towards a taxonomy of content sensitivity and sharing preferences for photos. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[16] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2, 1 (2020), 2522–5839.

[17] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

[19] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[20] Francesca Mosca and Jose Such. 2022. An explainable assistant for multiuser privacy. *Autonomous Agents and Multi-Agent Systems* 36, 1 (2022), 1–45.

[21] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2017. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE international conference on computer vision*. 3686–3695.

[22] Lloyd S Shapley. 1997. A value for n-person games. *Classics in game theory* 69 (1997).

[23] Anna Cinzia Squicciarini, Andrea Novelli, Dan Lin, Cornelia Caragea, and Haoti Zhong. 2017. From tag to protect: A tag-driven policy recommender system for image sharing. In *2017 15th Annual Conference on Privacy, Security and Trust (PST)*. IEEE, 337–33709.

[24] Ashwini Tonge and Cornelia Caragea. 2020. Image privacy prediction using deep neural networks. *ACM Transactions on the Web (TWEB)* 14, 2 (2020), 1–32.

[25] Onuralp Ulusoy and Pinar Yolum. 2021. PANOLA: A Personal Assistant for Supporting Users in Preserving Privacy. *ACM Transactions on Internet Technology (TOIT)* 22, 1 (2021), 1–32.

[26] Tony Vila, Rachel Greenstadt, and David Molnar. 2004. Why we can't be bothered to read privacy policies. In *Economics of information security*. Springer, 143–153.

[27] Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, and Elena Demidova. 2012. Privacy-aware image classification and search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 35–44.

[28] Chenye Zhao, Jasmine Mangat, Sujay Koujalgi, Anna Squicciarini, and Cornelia Caragea. 2022. PrivacyAlert: A Dataset for Image Privacy Prediction. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1352–1361.