Explain to Me: Towards Understanding Privacy Decisions

Extended Abstract

Gonul Ayci
Bogazici University
Istanbul, Turkey
gonul.ayci@boun.edu.tr

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

Murat Şensoy
Ozyegin University
Istanbul, Turkey
drmuratsensoy@gmail.com

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

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 work 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.

KEYWORDS

Privacy; explainability; online social networks

ACM Reference Format:
Gonul Ayci, Arzucan Özgür, Murat Şensoy, and Pınar Yolum. 2023. Explain to Me: Towards Understanding Privacy Decisions: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 2 pages.

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. 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. Privacy assistants that work side by side with humans in a decentralized manner could help them by suggesting what content is private [2]. However, for them to be successful, the privacy assistants need to be able to explain their privacy decision to their users.

Existing work on an explanation for binary classifications generally consider what features of the classification have been influential for the classification [3]. Even though these approaches are important because they provide interpretability of the underlying classifier, explanations would be difficult to understand when a model has a high-dimensional feature space. They are not meant to provide explanations to the end user, which we we aim for 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 relations 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.

2 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. 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, 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 explanation is augmented with a short description using a predetermined language structure to explain the visual representation and the relations between topics. The text is thus supplementary and does not provide additional information.

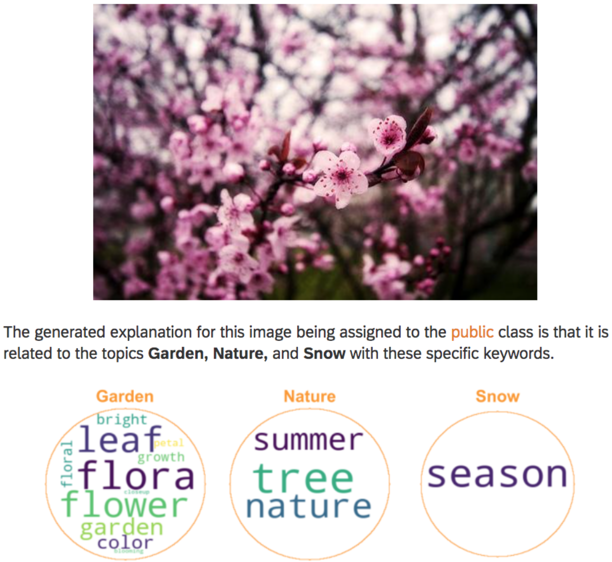
Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

In order to realize the above explanations, we need to understand how we can associate images with topics. 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. Our proposed method automatically generates descriptive keywords for each image using a tool. The method assigns weights to each keyword based on how relevant a keyword is to a given collection of keyword sets. Then, it explores topics present in a dataset of images using image-keyword relation by making use of a topic modelling technique. We named the topics manually, for example, Nature, Child, and Fashion. After building a model using a topic modeling technique, we use the Random Forest algorithm to make predictions.

The TreeExplainer [3] model is a specific implementation of the SHAP (SHapley Additive exPlanations) approach, which can be used to understand how a ML model arrived at its prediction. The TreeExplainer model computes the contribution of each feature to a prediction, taking into account the interactions between features using tree-based models such as the Random Forest algorithm. Not all features have an equal contribution to a class prediction: a feature can push the prediction higher (positive SHAP value) or lower (negative SHAP value), and their magnitude can differ. A ML model concludes its prediction by taking into account the contribution of each feature. This is useful in interpreting how the model 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 modify the TreeExplainer output by reducing the number of features shown to the user.

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 SHAP 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 SHAP values when creating the explanations. Our methodology generates human-understandable explanations through a topic reduction in the output of the TreeExplainer model. It can display a single topic that has a relatively high contribution compared with others. On the other hand, the generated explanation can display multiple topics that their contributions can be together or opposing forces on the prediction.

Figure 1 shows an example image, which is annotated as public, and the explanation for the image in the proposed explanation schema. The explanation displays three different topics namely Garden, Nature, and Snow with their related keywords that can help to explain why the given image has been annotated as public. Also, it provides a text that helps to understand the relation of these three topics. For the given image, each topic contributes significantly to making the prediction public.



The generated explanation for this image being assigned to the public class is that it is related to the topics **Garden, Nature,** and **Snow** with these specific keywords.

Figure 1: Example public image and its generated explanation

3 CONCLUSION

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 modeling techniques 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. An important method to evaluate the work will be to conduct a detailed user study to understand if the participants find the generated explanations sufficient, satisfying, and understandable. An interesting other dimension is to evaluate if these explanations can be useful in understanding the behavior of privacy assistants that are based on deep learning architectures [1].

4 ACKNOWLEDGMENTS

The first author is supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) and Turkish Directorate of Strategy and Budget under the TAM Project number $2007K12-873$. This research was partially funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl.

REFERENCES

[1] Gönül Aycı, Murat Şensoy, Arzucan Özgür, and Pınar Yolum. 2023. Uncertainty-aware Personal Assistant for Making Personalized Privacy Decisions. *ACM Transactions on Internet Technology (TOIT)* (2023). To appear.

[2] Nadin Kökciyan and Pınar Yolum. 2022. Taking Situation-Based Privacy Decisions: Privacy Assistants Working with Humans. In *International Joint Conference on AI (IJCAI)*. 703–709.

[3] 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.